OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

EIT

FAKULTÄT FÜR
ELEKTROTECHNIK UND
INFORMATIONSTECHNIK

Toward Accurate, Reliable and Efficient Gaze Estimation

Dissertation
zur Erlangung des akademischen Grades

Doktoringenieur
(Dr.-Ing.)

von **M.Sc. Ahmed Awadalla Ahmed Soliman Abdelrahman**
geb. am 17.10.1989 in Kalyobiya, Egypt

genehmigt durch die Fakultät für Elektrotechnik und Informationstechnik
der Otto-von-Guericke-Universität Magdeburg

Gutachter: Prof. Dr.-Ing. habil. Ayoub Al-Hamadi
Prof. Dr. Aly A. Farag
Prof. Dr. Ahmed Al-Dubai

eingereicht am 13.06.2024
Promotionskolloquium am 22.01.2025

Contents

List of Figures

List of Tables

Abstract

Gaze estimation has become increasingly vital across various fields, such as human-computer interaction, autonomous systems, and assistive technologies. It plays a fundamental role in interpreting human focus and intentions, improving user experience, promoting accessibility, and ensuring safety. By accurately determining the direction of gaze, interfaces and devices can become significantly more intuitive, enriching interactions between humans and machines. However, accurate gaze estimation faces challenges, especially in unconstrained settings, due to variability in lighting, head poses, facial expressions, and occlusions. Traditional methods struggle with real-world applicability, as they require specialized hardware and controlled environments. Deep learning, particularly Convolutional Neural Networks (CNNs), has improved gaze estimation accuracy by handling complex and high-dimensional data. However, robust gaze estimation requires careful consideration of network architecture, loss functions, and the learning process to manage the complexities of gaze datasets.

Despite advances in gaze estimation using CNNs, several significant challenges remain. Ensuring high accuracy is difficult due to the need to extract fine-grained gaze features from facial images, compounded by individual anatomical differences and subjective biases. Achieving reliability in cross-dataset evaluations is also challenging due to variability in datasets, the complexity of dataset collection, and annotation. Additionally, high computational costs pose a challenge, especially in contexts requiring real-time performance and multimodal data integration. This thesis aims to develop a gaze estimation model that addresses these challenges by achieving high balance between accuracy, reliability, and efficiency. This model should be able to estimate gaze accurately within dataset settings, maintain performance across diverse datasets, and operate in real-time using available computational resources effectively.

A significant contribution of this thesis is a comprehensive survey of both conventional and deep learning-based gaze estimation methods. The survey categorizes existing methods into conventional and deep learning-based approaches, spotlighting three principal conventional techniques: model-based, feature-based, and appearance-based. This survey critically examines the progression and efficacy of existing deep learning gaze estimation methods, identifying their strengths and limitations. It highlights the importance of validating these models against robust and diverse datasets that accurately reflect real-world conditions, and reviews critical benchmark datasets used to ensure model reliability and effectiveness across various environments.

Then, this thesis presents L2CS-Net to improve the accuracy of gaze estimation. It features a dual-branch CNN architecture that separately predicts horizontal and vertical gaze angles with a multi-loss approach, incorporating both classification and regression losses. This design allows for precise learning of discriminative features specific to each angle by separating prediction tasks into distinct fully connected layers. Moreover, the multi-loss approach optimizes the model to leverage the strengths of both classification for coarse gaze direction estimation and regression for fine-grained predictions, significantly enhancing overall accuracy.

Further, the proposed MTGH-Net improves the reliability of gaze estimation by integrating gaze and head pose estimation into a single multi-task learning framework. It tackles the gaze generalization challenge by employing advanced training approach that utilizes two separate datasets, one for gaze and the other for head pose. This method allows MTGH-Net to benefit from the increased data from these diverse datasets, leading to enhanced understanding of unseen data and more robust representations for both tasks. Furthermore, MTGH-Net introduces a simplified and efficient 6D-parameter rotation matrix representation coupled with a geodesic-based loss function for both gaze and head pose estimation tasks to overcome the discontinuity problem inherent in traditional gaze representation methods and ensuring the model's learning process is not biased toward either task.

Additionally, the development of MGAZE-Net introduces an innovative solution to balance performance with computational efficiency. This novel and lightweight CNN architecture is augmented with a progressive combination of attention mechanisms, including Squeeze-and-Excitation (SE), Convolutional Block Attention Module (CBAM), and Coordinate Attention (CA). These mechanisms are strategically designed to systematically emphasize important gaze information by capturing local and global spatial relationships within facial images. The strategic placement of these mechanisms allows MGAZE-Net to extract fine-grained features relevant to gaze estimation with remarkable efficiency, avoiding the computational overhead typically associated with deep CNN models and transformers.

All of the models proposed in this thesis undergo extensive evaluation and ablation studies, which have been shown to provide the best balance between performance and efficiency over state-of-the-art methods. These contributions collectively fulfill the goal of designing a model capable of estimating gaze with higher accuracy, reliability, and efficiency compared to existing methods, establishing a new benchmark in this field.

Deutsche Kurzfassung

Die Bestimmung der Blickrichtung wird in einer Vielzahl von Bereichen wie der Mensch-Computer-Interaktion, autonomen Systemen und unterstützenden Technologien zunehmend wichtig. Sie spielt eine grundlegende Rolle bei der Interpretation menschlicher Aufmerksamkeit und Absichten und verbessert so die Benutzererfahrung, fördert die Zugänglichkeit und gewährleistet die Sicherheit. Durch die genaue Bestimmung der Blickrichtung können Schnittstellen und Geräte wesentlich intuitiver gestaltet und die Interaktion zwischen Menschen und Maschinen erheblich bereichert werden. Eine exakte Schätzung der Blickrichtung ist jedoch eine Herausforderung, vor allem in unbeschränkten Umgebungen, da hier die Beleuchtung, Kopfhaltung, Mimik und verdeckte Bereiche variieren. Herkömmliche Methoden sind in der realen Welt nur schwer anwendbar, da sie spezielle Hardware und kontrollierte Umgebungen erfordern. Deep Learning, insbesondere Convolutional Neural Networks (CNNs), haben die Genauigkeit der Blickschätzung durch die Verarbeitung komplexer und hochdimensionaler Daten verbessert. Eine effektive Anwendung erfordert jedoch eine sorgfältige Prüfung der Netzwerkarchitektur, der Verlustfunktionen und des Lernprozesses, um die Komplexität der Ausgangsdaten der Blickrichtungsbestimmung zu verarbeiten.

Trotz der Fortschritte bei CNNs gibt es nach wie vor einige große Herausforderungen, darunter die Genauigkeit, Zuverlässigkeit und Effizienz der Bestimmung der Blickrichtung. Die Gewährleistung einer hohen Genauigkeit ist schwierig, da detaillierte Blickmerkmale aus Gesichtsbildern extrahiert werden müssen, was durch individuelle anatomische Unterschiede und subjektive Verzerrungen erschwert wird. Zuverlässigkeit in der Evaluierung über verschiedene Datensätze hinweg zu erreichen ist aufgrund der Variabilität der Datensätze, der Komplexität der Datensatzerfassung und -annotation sowie des Zusammenspiels von Augen- und Kopfbewegungen eine Herausforderung. Darüber hinaus stellt der hohe Rechenaufwand eine Herausforderung dar, vor allem in Kontexten, die Echtzeitleistung und multimodale Datenintegration erfordern. Ziel dieser Arbeit ist es, ein Modell zur Schätzung der Blickrichtung zu entwickeln, das diesen Herausforderungen durch eine hohe Balance von Genauigkeit, Zuverlässigkeit und Effizienz begegnet. Das Modell soll in der Lage sein, die Blickrichtung innerhalb von Datensatzumgebungen genau zu schätzen, die Leistung über verschiedene Datensätze hinweg beizubehalten und in Echtzeit mit den verfügbaren Rechenressourcen effektiv zu arbeiten.

Ein wesentlicher Beitrag dieser Arbeit ist ein umfassender Überblick über konventionelle und Deep-Learning-basierte Blickrichtungsschätzungsmethoden. Die Übersicht kategorisiert die existierenden Methoden in konventionelle und Deep-Learning-basierte

Ansätze, wobei die drei wichtigsten konventionellen Techniken hervorgehoben werden: modellbasierte, merkmalsbasierte und erscheinungsbasierte Methoden. Die Studie untersucht kritisch die Entwicklung und Wirksamkeit bestehender Deep-Learning-Methoden zur Blickrichtungsschätzung und identifiziert deren Stärken und Grenzen. Es wird hervorgehoben, wie wichtig es ist, diese Modelle anhand von robusten und vielfältigen Datensätzen zu validieren, die die realen Bedingungen genau widerspiegeln, und es werden kritische Benchmark-Datensätze überprüft, die verwendet werden, um die Zuverlässigkeit und Effektivität des Modells in verschiedenen Umgebungen sicherzustellen.

In dieser Arbeit wird das L2CS-Net zur Verbesserung der Genauigkeit der Blickrichtungsschätzung vorgestellt. Es verfügt über eine CNN-Architektur mit zwei Verzweigungen, die horizontale und vertikale Blickwinkel separat mit einem Multi-Loss-Ansatz vorhersagt, der sowohl Klassifikations- als auch Regressionsverluste beinhaltet. Dieses Design ermöglicht ein präzises Lernen von Unterscheidungsmerkmalen, die für jeden Winkel spezifisch sind, indem die Vorhersageaufgaben in verschiedene voll verbundene Schichten aufgeteilt werden. Darüber hinaus optimiert der Multi-Loss-Ansatz das Modell, um die Stärken sowohl der Klassifikation für die grobe Schätzung der Blickrichtung als auch der Regression für detaillierte Vorhersagen zu nutzen, was die Gesamtgenauigkeit deutlich erhöht.

Darüber hinaus verbessert das vorgeschlagene MTGH-Netz die Zuverlässigkeit der Blickrichtungsschätzung durch die Integration von Blickrichtungs- und Kopfposenschätzung in einem einzigen Multi-Task-Lernsystem. Der Herausforderung der Generalisierung der Blickrichtung wird durch einen fortschrittlichen Trainingsansatz begegnet, indem zwei separate Datensätze verwendet werden, einer für die Blickrichtung und der andere für die Kopfhaltung. Diese Methode ermöglicht es dem MTGH-Net, von der größeren Anzahl an Daten aus diesen unterschiedlichen Datensätzen zu profitieren, was zu einem besseren Verständnis von ungesehenen Daten und robusteren Repräsentationen für beide Aufgaben führt. Darüber hinaus führt das MTGH-Net eine vereinfachte und effiziente 6D-Parameter-Rotationsmatrix-Darstellung ein. Diese ist mit einer geodätisch basierten Verlustfunktion sowohl für Blickrichtung- als auch für Kopfposenschätzungsaufgaben gekoppelt, um das Diskontinuitätsproblem zu überwinden und um sicherzustellen, dass der Lernprozess des Modells nicht für eine der beiden Aufgaben voreingenommen ist.

Zudem stellt die Entwicklung von MGAZE-Net eine innovative Lösung dar, die ein Gleichgewicht zwischen Leistung und Recheneffizienz schafft. Diese neue und schlanke CNN-Architektur wird durch eine fortschrittliche Kombination von Aufmerksamkeitsmechanismen ergänzt, darunter Squeeze-and-Excitation (SE), Convolutional Block Attention Module (CBAM) und Coordinate Attention (CA). Diese Mechanismen sind strategisch so konzipiert, dass sie systematisch wichtige Informationen zur Blickrichtung hervorheben, indem sie sowohl lokale als auch globale räumliche Beziehungen innerhalb von Gesichtsbildern erfassen. Die strategische Platzierung dieser Mechanismen ermöglicht es MGAZE-Net, detaillierte Merkmale, die für die Abschätzung der Blickrichtung relevant sind, mit bemerkenswerter Effizienz zu extrahieren und dabei zusätzlichen Rechenaufwand zu vermeiden, der typischerweise mit tiefen CNN-Modellen und Transformatoren verbunden ist.

Alle in dieser Arbeit vorgestellten Modelle wurden umfangreichen Evaluierungs- und Ablationsstudien unterzogen, wobei sich herausstellte, dass sie ein optimales Gleichgewicht zwischen Leistung und Effizienz im Vergleich zu Methoden auf dem aktuellen Stand der Technik bieten. Gemeinsam erreichen diese Beiträge erfolgreich das Ziel, ein Modell zu entwickeln, das in der Lage ist, Blickrichtungen mit optimaler Genauigkeit, Zuverlässigkeit und Effizienz abzuschätzen.

Related Publications

Most of the material contained in this dissertation is partly based on the following refereed papers and journals published in a variety of peer-reviewed journals and international conference proceedings.

Peer-reviewed Articles in International Journals & Conferences:

[1] A. Abdelrahman, T. Hempel, A. Khalifa, and A. Al-Hamadi, "L2CS-Net: Fine-grained gaze estimation in unconstrained environments," *In 2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*, pp. 98-102. IEEE, 2023.

[2] A. Abdelrahman, D. Strazdas,A. Khalifa, J. Hintz, T. Hempel, and A. Al-Hamadi, "Multimodal engagement prediction in multiperson human–robot interaction," *IEEE Access, **(IF 3.9)***, 10:61980–61991, 2022.

[3] T. Hempel, A. Abdelrahman, and A. Al-Hamadi, "Toward Robust and Unconstrained Full Range of Rotation Head Pose Estimation," *IEEE Transactions on Image Processing, **(IF 12)***, 33 (2024): 2377-2387.

[4] T. Hempel, A. Abdelrahman, and A. Al-Hamadi, "6d rotation representation for unconstrained head pose estimation," *In 2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2496-2500. IEEE, 2022.

[5] A. Khalifa, A. A. Abdelrahman, D. Strazdas, J. Hintz, T. Hempel, and A. Al-Hamadi, "Face recognition and tracking framework for human-robot interaction," *Applied Sciences, **(IF 2.9)***, 12(11):5568, 2022.

[6] A. A. Abdelrahman, T. Hempel, and A. Al-Hamadi, "MTGH : Multi-task Gaze and Head Pose Estimation in-the-Wild," *Neural Computing and Applications, **(IF 6)***, Under Review.

[7] A. A. Abdelrahman, T. Hempel, A. Khalifa, and A. Al-Hamadi, "Fine-grained gaze estimation based on the combination of regression and classification losses," *Applied Intelligence, **(IF 5.3)***, Under Review.

[8] A. A. Abdelrahman, T. Hempel, A. Khalifa, and A. Al-Hamadi, "MobGazeNet : Robust Gaze Estimation Mobile Network Based On Progressive Attention Mechanisms.," *Machine Vision and Applications, **(IF3.3)***, Under Review.

[9] A. A. Abdelrahman, B. Al-Tawil, and A. Al-Hamadi, "Deep Learning Based Gaze Estimation: A Review," *Sensors, **(IF 2.9)***, Under Review.

[10] D. Strazdas, J. Hintz, A. Khalifa, A. A. Abdelrahman, T. Hempel, and A. Al-Hamadi, "Robot system assistant (rosa): Towards intuitive multi-modal and multi-device human-robot interaction," *Sensors,* ***(IF 3.9)***, 22(3):923, 2022.

[11] A. Khalifa, A. A. Abdelrahman, T. Hempel, and A. Al-Hamadi, "Towards efficient and robust face recognition through attention-integrated multi-level cnn," *Multimedia Tools and Applications,* ***(IF 3.6)***, Under Review.

[12] T. Hempel, M Jung, A. Abdelrahman, and A. Al-Hamadi, "NITEC: Versatile Hand-Annotated Eye Contact Dataset for Ego-Vision Interaction," *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4437-4446. 2024.

[13] B. Al-Tawil, T Hempel, A.A Abdelrahman, and A. Al-Hamadi, "A review of visual SLAM for robotics: evolution, properties, and future applications" *Frontiers in Robotics and AI ,* ***(IF 3)***, 1347985.

[14] L. Dinges, M.-A. Fiedler, A. Al-Hamadi, A. Abdelrahman, J Weimann, and D Bershadskyy, "Uncovering lies: deception detection in a rolling-dice experiment," *In International Conference on Image Analysis and Processing,* , pp. 293-303. 2024.

[15] L. Dinges, M.-A. Fiedler, A. Al-Hamadi, T. Hempel, A. Abdelrahman, J Weimann, and D Bershadskyy, "Automated Deception Detection from Videos: Using End-to-End Learning Based High-Level Features and Classification Approaches," *Neural Computing and Applications,* ***(IF 6)***, Under Review.

1 Introduction

Human gaze estimation has gained considerable interest in the field of computer vision, given its pivotal role in understanding human attention, intention, and interaction. The gaze direction offers invaluable insights into a person's focus of attention, making it a crucial component in diverse applications ranging from human-robot interaction to autonomous driving and from augmented reality to assistive technologies. The ability to accurately estimate human gaze can make interactions between humans and machines more intuitive and efficient. However, accurate gaze estimation, especially in unconstrained or in the wild settings, remains a big challenge. Conventional approaches require specialized hardware and are often limited to controlled laboratory settings, limiting their applicability in real-world scenarios. Variability in lighting conditions, head poses, facial expressions, and occlusions further increase the challenge, making it difficult to achieve high accuracy with conventional methods.

The advent of deep learning, particularly Convolutional Neural Networks (CNNs), has revolutionized the field of gaze estimation. These models have shown an exceptional capability to handle the complex, high-dimensional data features of real-world images, paving the way for more accurate and robust gaze estimation methods. However, effective application of CNNs in gaze estimation requires careful consideration of the network architecture, loss functions, and the learning process to accommodate the complex nature of gaze data. In unconstrained environments, the gaze estimation problem is compounded by the inherent variability and unpredictability of natural settings. The small size of the eye region relative to the entire face, significant variations in head pose, and various lighting conditions present substantial obstacles. These factors require innovative approaches that can adapt and learn from the complexity and diversity of real-world data. This thesis addresses these challenges by exploring novel deep learning architectures and loss functions specifically designed for gaze estimation in unconstrained environments. Through the development of two-branch CNN architectures that separately predict gaze angles and the introduction of a multi-loss approach that combines classification and regression objectives, this work aims to push the boundaries of accuracy and efficiency in gaze estimation. By leveraging the power of deep learning and the richness of large-scale, unconstrained gaze datasets, the research presented in this thesis contributes to the advancement of gaze estimation technology, making it more adaptable and effective for real-world applications.

The importance of this research lies not only in its technical contributions but also in its potential to facilitate more natural and intuitive human-machine interactions. As technology continues to evolve, the ability to accurately interpret human gaze

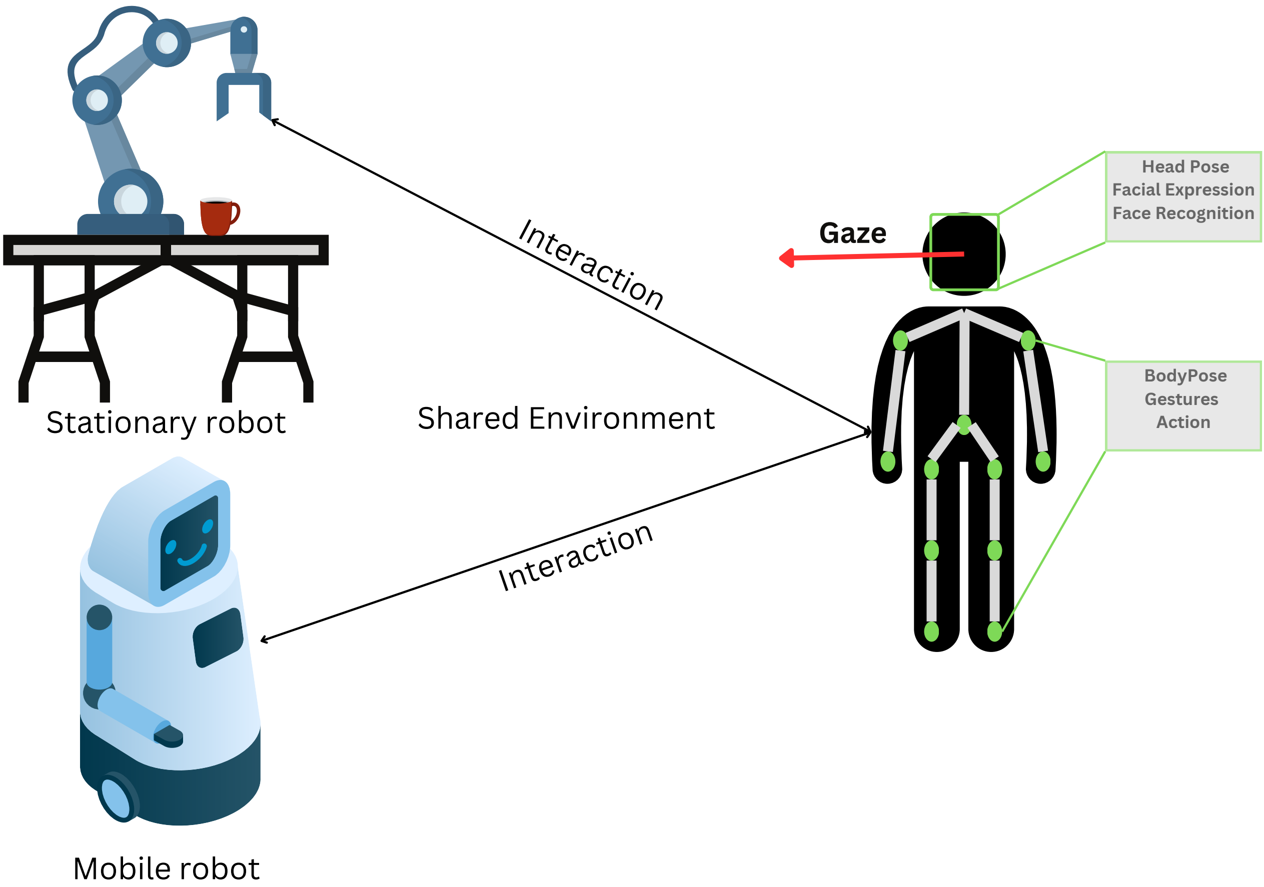


Figure 1.1: A major factor affecting the accuracy of gaze estimation within datasets is the challenge of extracting fine-grained gaze features from the small eye region in facial images, compounded by subjective bias.

will play a crucial role in bridging the gap between human intentions and machine understanding.

1.1 Motivation

Human gaze estimation is employed in various applications across multiple domains, demonstrating its versatility and importance in modern technological contexts. Each of these applications leverages the fundamental capability of gaze estimation technology to infer a person's focus of attention and intentions from the direction of their gaze, opening up new possibilities for interaction and understanding across a wide range of domains.

1.1.1 Enhancing Human-Robot Interaction

Human-Robot Interaction (HRI) relies on various non-verbal cues to facilitate natural and effective communication. These cues include head pose, facial expression, body pose and gestures. These cues are crucial for robots as they help interpret human attention, emotions, and intentions and to enable more natural and intuitive interactions. Among these, gaze is a fundamental non-verbal cue that helps establish understanding, mutual awareness, and a common ground for interaction. For robots to interact effectively with humans, they must be able to understand and respond to human social cues, with gaze direction being one of the most critical. Gaze estimation allows robots to infer what a person is paying attention to, enhancing the robot's ability to engage in meaningful interactions.

Social eye gaze is particularly important because it is closely tied to what people are thinking and doing, making it a powerful indicator of attention and intention. Eye gaze can be used to convey information, regulate social intimacy, manage turn-taking, and convey social or emotional states. For example, mutual gaze, often referred to as "eye contact," establishes a connection between the robot and the human, enhancing engagement and trust. Referential gaze, or gaze directed at an object, helps robots guide human attention to specific items or areas, facilitating tasks such as collaborative work or guided tours.

In educational settings, robots that can follow and respond to human gaze can better understand and adapt to student needs, making learning more interactive and personalized. In healthcare, robots that can interpret gaze can better assist patients, particularly those with communication impairments, by understanding their needs and intentions without verbal communication.

Furthermore, incorporating gaze estimation into robots enhances their social presence and naturalness, making them more effective in roles that require human-like interactions. By accurately estimating and responding to human gaze, robots can perform tasks more efficiently and effectively, whether they are serving as companions, assistants, or team members in various domains.

1.1.2 Gaze Applications In Various Fields

Human-Computer Interaction (HCI): Gaze estimation is revolutionizing HCI by enabling computers and devices to understand the user's point of focus. This facilitates more intuitive interfaces, where actions can be controlled by eye movements, enhancing accessibility for users with mobility impairments.

Augmented Reality (AR) and Virtual Reality (VR): In AR and VR, gaze estimation can improve the realism and interactivity of the experience. By tracking where a user is looking, content can be dynamically adjusted, ensuring that computational resources are focused on rendering the parts of the scene the user is actually observing. This leads to more immersive experiences with better allocation of processing power.

Autonomous Driving: Gaze estimation can enhance safety features in automotive technology by monitoring the driver's gaze direction to assess their level of attention. For autonomous vehicles, understanding the gaze direction of pedestrians and other drivers can improve decision-making processes and predict their actions, enhancing safety for all road users.

Healthcare: Gaze estimation finds applications in healthcare, particularly in the diagnosis and monitoring of certain conditions. For example, tracking eye movements can help diagnose neurological disorders, assess visual impairments, or monitor the effectiveness of treatments. In addition, gaze-controlled interfaces can assist people with physical disabilities, enabling them to interact with computers and other devices using eye movements.

Assistive Technologies: Gaze estimation provides a communication pathway for individuals with severe physical disabilities, enabling them to interact with computers and their environment using their eyes as input devices. This can include typing, operating home appliances or navigating wheelchairs.

Educational Tools: In educational settings, gaze estimation can help in understanding student engagement and focus. By analyzing where a student is looking, educators can assess the effectiveness of teaching materials and methods, adjusting them in real time or for future lessons to improve learning outcomes.

Psychological and Marketing Research: Gaze estimation offers valuable insights into consumer behavior by revealing what catches their attention in advertisements, stores, or online platforms. This data can inform product placement, store layout designs, and advertising strategies to maximize engagement and sales.

Security and Surveillance: In security contexts, gaze estimation can be used to identify suspicious behaviors by tracking where individuals are looking, especially in crowded or sensitive environments. This can help in the early detection of potential threats.

Social Robotics and Human-Robot Interaction: For robots to interact effectively with humans, understanding human attention and intentions is crucial. Gaze estimation allows robots to perceive where a human is looking, facilitating more natural and intuitive interactions.

The application of gaze estimation technology across different fields comes with specific requirements for accuracy and efficiency, leading to a need for tailored algorithmic approaches. High accuracy is crucial for applications like healthcare, where accurate gaze estimation can contribute to diagnostic processes and assistive technologies, and patient monitoring systems. In contrast, applications that require real-time interac-

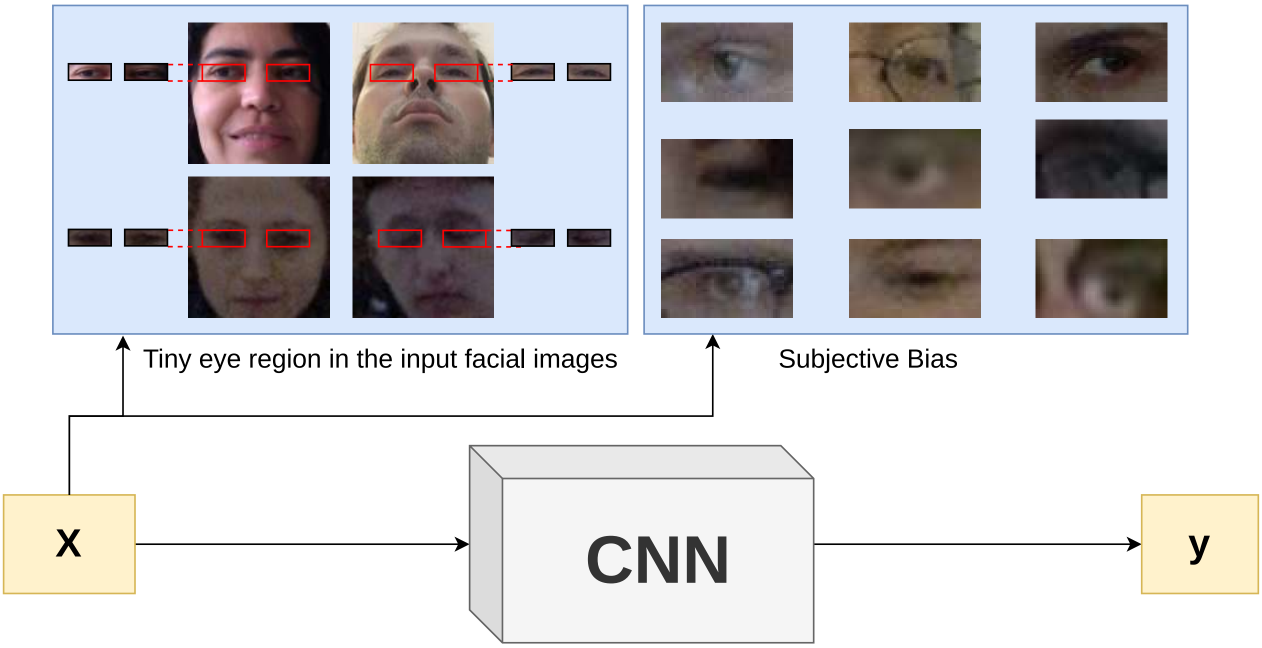


Figure 1.2: A major factor affecting the accuracy of gaze estimation within datasets is the challenge of extracting fine-grained gaze features from the small eye region in facial images, compounded by subjective bias.

tion, such as gaze-based interfaces in gaming or VR, require high frame rates for a seamless user experience. The choice of gaze estimation algorithms depends on these requirements, often necessitating a trade-off between efficiency and accuracy. The challenge of optimizing these critical factors, accuracy and efficiency, according to the specific demands of each application not only underscores the complexity of this field but also serves as a strong motivation for innovation.

1.2 Challenges

In the field of gaze estimation from facial images, the task of training robust gaze estimation models is notably affected by three major challenges include **accuracy**, **reliability** and **efficiency**. This section delves into these challenges, highlighting the intricate processes that currently hinder the development of robust gaze estimation models.

C1. Accuracy: One of the primary challenges in gaze estimation is ensuring high accuracy in within-dataset evaluations. This challenge arises from the difficulty of extracting fine-grained gaze features from facial images, compounded by subjective bias due to individual differences in eye anatomy, such as variations in the nodal point, iris size and color, eyelid shape and eye openness. Accurate estimation of gaze direction from facial images involves two critical tasks: (a) capturing rich and fine-grained features from the appearance of the eye area and (b) filtering out irrelevant features of

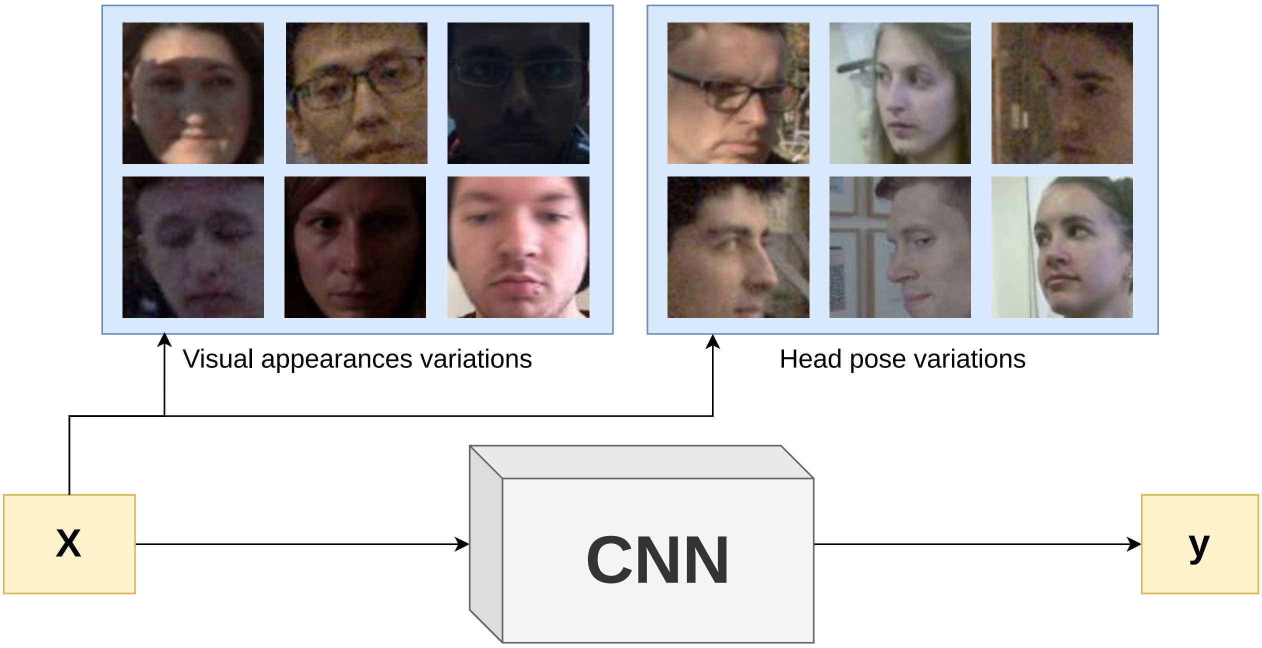


Figure 1.3: Major factors affecting the reliability of gaze estimation in cross-dataset evaluations include variations in visual appearance and head poses.

gaze from facial images. Complexity increases as the eye region is only a small part of the facial image and varies greatly from person to person, which affects the accuracy of the models. To overcome these challenges, various CNNs have been developed with common backbones such as LeNet [248], Alex-Net [107], VGG [54], ResNet-18 [95], ResNet-50 [242], or they rely on specially designed CNN architectures [30, 36, 113, 230] to extract fine-grained gaze features from facial images. However, CNNs face the challenge of compensating for subjective biases as they tend to focus on large and continuous areas, making it difficult to extract rich features corresponding to the appearance of the eyes. **Given this context**, several strategies and innovations can be used to enhance the ability of CNNs to capture fine-grained features of eye appearance for a more accurate estimation of gaze.

C2. Reliability: Another significant challenge is to achieve high performance in cross-dataset evaluations. This refers to the model's ability to maintain high accuracy when applied to unseen domains. This challenge is compounded by several issues:

- Variability in existing datasets: Existing datasets [54, 55, 86, 107, 168, 171, 252] often lack size and diversity, which introduces biases that degrade the performance of cross-dataset evaluations. These datasets are collected using varied cameras and setups, leading to significant variations in visual appearances, such as differences in resolution, lighting conditions, and head pose values. This diversity in data collection methodologies further complicates the generalization capabilities of gaze estimation models.

- Dataset collection and annotation: creating large-scale datasets that capture a wide range of gaze values across different populations (age groups, ethnicities, etc.) and in various settings (indoor, outdoor, etc.) is both time-consuming and costly. Furthermore, accurately annotating these datasets with a precise 3D gaze direction is fraught with potential errors. This can be due to technical limitations in estimating the positions of eyes and gaze targets, as well as human factors such as participant distraction or involuntary movements. These challenges constitute major barriers to developing robust gaze estimation technologies, limiting their generalizability and practical utility.
- Eye and Head Interplay: The complex relationship between eye and head movements is pivotal in accurately estimating gaze direction. This is because gaze direction is not solely determined by eye position but also by the orientation of the head. Current research approaches vary, with some models [107, 244] incorporating head pose implicitly through learning from large datasets, while others [54, 249] use it as an explicit feature to enhance model accuracy.

Therefore, the exploration of the relationship between eye and head in gaze estimation presents an opportunity for enhancing the accuracy and reliability of these systems. As the field progresses, there are several avenues and considerations for future research and development to further leverage this interplay.

C3. Efficiency: The challenge of high computational costs in gaze estimation is significant, especially in contexts where resources are limited or real-time performance is crucial. Deep learning-based gaze estimation methods require substantial computational power to accurately predict gaze direction in real-time. These models employ deep CNNs to extract subtle patterns in eye movements, which significantly increase the computational load. Moreover, most gaze estimation applications are multimodal, integrating multi input features along with gaze. For example, head pose is combined with gaze in various applications such as human-robot interaction (HRI), augmented reality, and autonomous driving. This multimodal integration presents even greater challenges in achieving accurate and efficient gaze estimation in these specific applications, due to the increased complexity and heightened data processing demands. To address this challenge, various network architectures are proposed to balance the accuracy required for gaze estimation with computational efficiency. **Therefore,** Achieving accurate, reliable and computationally efficient gaze estimation in various applications requires a trade-off between accuracy, reliability and computational cost.

1.3 Goal and Contributions

As illustrated in Fig. 4.1, the objective of this thesis is to design a model capable of estimating gaze with high accuracy, reliability, and efficiency. accuracy means reach-

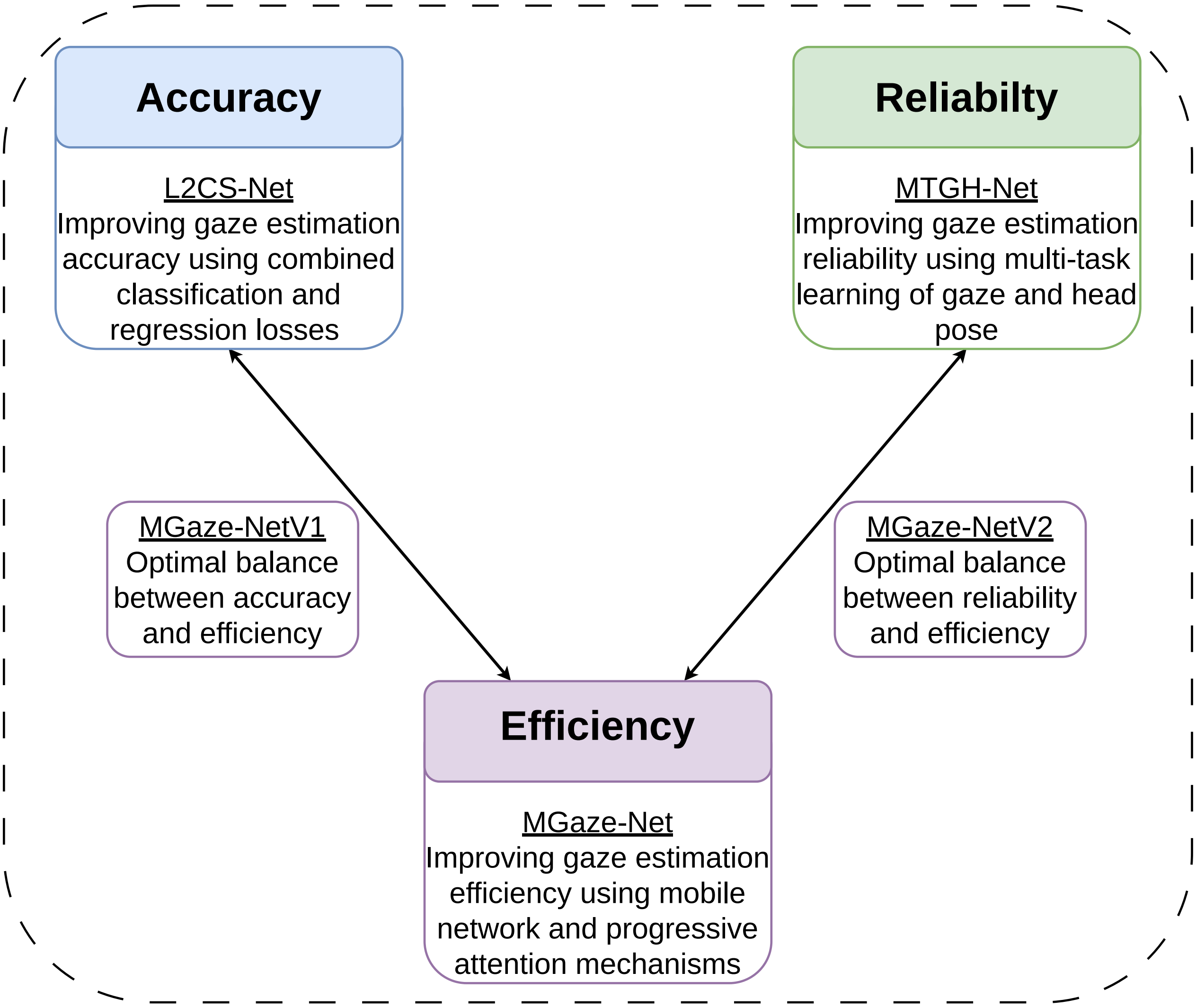

Accuracy
L2CS-Net
Improving gaze estimation accuracy using combined classification and regression losses
Reliabilty
MTGH-Net
Improving gaze estimation reliability using multi-task learning of gaze and head pose
MGaze-NetV1
Optimal balance between accuracy and efficiency
MGaze-NetV2
Optimal balance between reliability and efficiency
Efficiency
MGaze-Net
Improving gaze estimation efficiency using mobile network and progressive attention mechanisms

Figure 1.4: The goal of this thesis is to design a gaze estimation model that can estimate gaze accurately, reliably, and efficiently.

ing the accuracy necessary for within dataset settings. Reliable denotes the ability of the model to maintain high performance in diverse datasets and under different constraints. Efficiently means the model work in real-time, using resources effectively as needed. Aligned with this goal, the thesis introduces several significant contributions that address the complexities of gaze estimation. These key contributions are detailed below.

1. **Deep Learning Based Gaze Estimation Survey (*Survey*):** A comprehensive literature survey has been conducted to delve into recent advancements in the gaze estimation field (see Chapter 2). This survey categorizes and summarizes existing gaze estimation methods into conventional and deep learning-based approaches. It outlines three principal categories of conventional gaze estimation techniques: model-based, feature-based, and appearance-based, each described with their historical context, technological advancements, and inherent limita-

tions. As the narrative shifts focus to the impact of deep learning, it explores this domain from three perspectives: methodology, validation, and adaptation. The discussion delves into various architectural designs and learning strategies employed in deep learning models, highlighting the progression from simple convolutional networks to complex architectures involving recurrent networks and attention mechanisms that cater specifically to the nuanced needs of gaze estimation. It then addresses how these models are validated, emphasizing the importance of robust and diverse datasets that reflect a wide range of real-world conditions. Several benchmark datasets used in the field are reviewed, evaluating the performance metrics and experimental setups that help ensure the models' effectiveness and reliability across different environments and populations. Lastly, the focus shifts to adaptation techniques that enable deep learning models to perform well under varied operational conditions and to adapt to new users without extensive retraining. Techniques such as transfer learning, domain adaptation, and few-shot learning are discussed, underscoring their critical role in enhancing the flexibility and applicability of gaze estimation models.

2. **Improving Gaze Estimation Accuracy Using Combined Classification And Regression Losses:** a novel two-branch CNN architecture (see Chapter 3), designed to separately predict pitch and yaw gaze angles with a multi-loss approach incorporating both classification and regression losses. L2CS-Net can focus more effectively on learning discriminative features specific to each angle by separating the prediction of each gaze angle into distinct fully connected layers. Furthermore, the multi-loss approach enables the model to leverage the strengths of both classification for coarse gaze direction estimation and regression for fine-grained predictions, thereby enhancing the overall accuracy of gaze estimation. Demonstrated through extensive evaluation on four challenging datasets: MPIIFaceGaze, GazeCapture Gaze360, and RT-GENE L2CS-Net achieve state-of-the-art gaze accuracy with up to 18% improvement compared to the best performing methods. This shows its superior capability to accurately estimate the direction of the gaze in different environments. Comprehensive ablation study further validate the effectiveness of the proposed network architecture and loss function, underlining the significance of individual gaze angle prediction and the synergistic benefit of combining classification and regression losses. This approach addresses the challenge **C1** of gaze estimation by improving fine grained gaze features from the eye area, thus offering a robust and efficient solution for accurate gaze estimation.

3. **Improving Gaze Estimation Reliability Using Multi-task Learning:** An innovative framework (see Chapter 4) that leverage the intrinsic relationship between gaze and head pose estimation tasks through a multi-task learning paradigm. By integrating these tasks within a single network, MTGH-Net signif-

icantly reduces computational overhead and achieves synergistic improvements in estimation performance for both gaze and head pose. Further, MTGH-Net addresses the gaze generalization challenge by utilizing a new training approach that involves utilizing two separate datasets, one for gaze and the other for head pose. This approach allows MTGH-Net to benefit from the increased amount of data from the two diverse datasets, leading to a better understanding of unseen data and more robust representations for both tasks. This strategy enriches the model's exposure to large scale data with diverse environmental constraints, significantly enhancing its ability to generalize across datasets with varying features. This leads to an improvement in knowledge sharing and overall performance. In addition, MTGH-Net introduces a simplified and efficient 6D-parameter rotation matrix representation for both gaze and head pose estimation tasks. This representation, coupled with a geodesic-based loss function, enables accurate and direct regression of these tasks. The approach effectively overcomes the discontinuity problem inherent in traditional gaze representation methods and ensures that the model's learning process is not biased toward either task. The detailed evaluation and benchmarking of the MTGH-Net on four challenging datasets highlight its effectiveness in providing reliable gaze estimation. Further, MTGH-Net achieves state-of-the-art gaze reliability with up to a 27% improvement compared to the best reported results. This approach addresses the challenge **C2** of gaze estimation by training MTGH-Net on a large amount of data from two diverse datasets, enhancing its understanding of unseen data and strengthening its ability to generalize across various conditions and datasets.

4. **Improving Gaze Estimation Efficiency Using Mobile Network and Progressive Attention Mechanisms:** A novel and lightweight CNN architecture (see Chapter 5) augmented with a progressive combination of attention mechanisms, including Squeeze-and-Excitation (SE), Convolutional Block Attention Module (CBAM), and Coordinate Attention (CA). This hierarchical integration of attention mechanisms is designed to systematically emphasize important gaze information by capturing both local and global spatial relationships within facial images. The strategic placement of these mechanisms allows MGAZE-Net to extract fine-grained features relevant to gaze estimation with remarkable efficiency, without the computational overhead typically associated with deep CNN models and transformers. MGAZE-Net uses a rotation matrix formalism to represent gaze direction to mitigates the problems of discontinuity and ambiguity with spherical angle representation. MGAZE-Net utilizes a geodesic loss function which utilizes the geometric properties of rotation matrices and provides a more accurate measure of the discrepancy between predicted and ground truth gaze values. The efficacy and robustness of the proposed MGAZE-Net are rigorously validated through extensive experiments on four challenging datasets. The com-

prehensive evaluation demonstrates that MGAZE-Net, while utilizing the lowest computational resources, it outperforms existing state-of-the-art methods by up to 13% in accuracy and up to 25% in reliability. With these results, MGAZE-Net introduces a gaze estimation model that has the highest accuracy, reliability, and efficiency. The inclusion of an extensive ablation study within the MTGH-Net evaluation provides insightful analyses into the impact of its various components on overall performance. This approach addresses the three challenges **C1, C2 and C3** of gaze estimation by adapting three strong attention mechanisms to encode rich information for the eye appearance, while using mobile network architecture to improve efficiency.

In essence, by integrating the key contributions, an accurate, reliable and efficient gaze estimation model is presented and can be used in various applications. This streamlined model addresses the challenges inherent in real-world gaze estimation, especially within HRI settings.

1.4 Publications

This section comprises publications that have been authored during the progression of this Ph.D. thesis.

1.4.1 Thesis-related Publications

[1] A. Abdelrahman, T. Hempel, A. Khalifa, and A. Al-Hamadi, “L2CS-Net: Fine-grained gaze estimation in unconstrained environments,” *In 2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*, pp. 98-102. IEEE, 2023.

[2] A. Abdelrahman, D. Strazdas,A. Khalifa, J. Hintz, T. Hempel, and A. Al-Hamadi, “Multimodal engagement prediction in multiperson human–robot interaction,” *IEEE Access, **(IF 3.9)***, 10:61980–61991, 2022.

[3] T. Hempel, A. Abdelrahman, and A. Al-Hamadi, “Toward Robust and Unconstrained Full Range of Rotation Head Pose Estimation,” *IEEE Transactions on Image Processing, **(IF 12)***, 33 (2024): 2377-2387.

[4] T. Hempel, A. Abdelrahman, and A. Al-Hamadi, “6d rotation representation for unconstrained head pose estimation,” *In 2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2496-2500. IEEE, 2022.

[5] A. Khalifa, A. A. Abdelrahman, D. Strazdas, J. Hintz, T. Hempel, and A. Al-Hamadi, “Face recognition and tracking framework for human-robot interaction,” *Applied Sciences, **(IF 2.9)***, 12(11):5568, 2022.

[6] A. A. Abdelrahman, T. Hempel, and A. Al-Hamadi, “MTGH : Multi-task Gaze and Head Pose Estimation in the wild,” *Neural Computing and Applications,* ***(IF 6)***, Under Review.

[7] A. A. Abdelrahman, T. Hempel, A. Khalifa, and A. Al-Hamadi, “Fine-grained gaze estimation based on the combination of regression and classification losses,” *Applied Intelligence,* ***(IF 5.3)***, Under Review.

[8] A. A. Abdelrahman, T. Hempel, A. Khalifa, and A. Al-Hamadi, “MobGazeNet : Robust Gaze Estimation Mobile Network Based On Progressive Attention Mechanisms.,” *Machine Vision and Applications,* ***(IF3.3)***, Under Review.

[9] A. A. Abdelrahman, B. Al-Tawil, and A. Al-Hamadi, “Deep Learning Based Gaze Estimation: A Review,” *Sensors,* ***(IF 2.9)***, Under preparation.

[10] D. Strazdas, J. Hintz, A. Khalifa, A. A. Abdelrahman, T. Hempel, and A. Al-Hamadi, “Robot system assistant (rosa): Towards intuitive multi-modal and multi-device human-robot interaction,” *Sensors,* ***(IF 3.9)***, 22(3):923, 2022.

[11] A. Khalifa, A. A. Abdelrahman, T. Hempel, and A. Al-Hamadi, “Towards efficient and robust face recognition through attention-integrated multi-level cnn,” *Multimedia Tools and Applications,* ***(IF 3.6)***, Under Review.

[12] T. Hempel, M Jung, A. Abdelrahman, and A. Al-Hamadi, “NITEC: Versatile Hand-Annotated Eye Contact Dataset for Ego-Vision Interaction,” *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4437-4446. 2024.

[13] B. Al-Tawil, T Hempel, A.A Abdelrahman, and A. Al-Hamadi, “A review of visual SLAM for robotics: evolution, properties, and future applications” *Frontiers in Robotics and AI ,* ***(IF 3)***, 1347985.

[14] L. Dinges, M.-A. Fiedler, A. Al-Hamadi, A. Abdelrahman, J Weimann, and D Bershadskyy, “Uncovering lies: deception detection in a rolling-dice experiment,” *In International Conference on Image Analysis and Processing,* , pp. 293-303. 2024.

[15] L. Dinges, M.-A. Fiedler, A. Al-Hamadi, T. Hempel, A. Abdelrahman, J Weimann, and D Bershadskyy, “Automated Deception Detection from Videos: Using End-to-End Learning Based High-Level Features and Classification Approaches,” *Neural Computing and Applications,* ***(IF 6)***, Under Review.

1.5 Outline

The thesis is structured as follows.

- **Chapter 2** reviews existing methods related to the thesis, including conventional-based gaze estimation and deep learning-based gaze estimation. The representatives of each category are presented in detail.

- **Chapter 3** presents L2CS-Net, a dual branch multi-loss CNN-based network designed for the direct estimation of gaze direction from facial images. The strategy of predicting each gaze angle pitch and yaw through dedicated fully connected layers is pivotal for isolating and capturing the nuanced features specific to each angle's subspace, thereby enhancing the model's accuracy. To refine the accuracy of gaze predictions, a novel gaze multi-loss function is proposed for each gaze angle. This function combines regression and classification losses, facilitating a joint optimization process that significantly enriches the feature extraction process. By doing so, it becomes possible to distill more informative gaze features, substantially elevating the overall accuracy of gaze estimation.

- **Chapter 4** introduces MTGH-Net, a multi task gaze and head pose estimation framework designed to harness the intrinsic correlation between gaze and head pose. It encompass an advanced training strategy capable of integrating two distinct datasets, each with its own set of annotations for gaze and head pose, effectively minimizing the impact of variations in appearance and head pose. This strategy enriches the model's ability to generalize gaze estimation across diverse settings by distilling knowledge from images under various conditions.

- **Chapter 5** presents MGAZE-Net, a novel gaze estimation mobile network designed to tackle the challenges posed by capturing fine-grained gaze features in the constrained eye region of face images. By progressively combining attention mechanisms SE, CBAM and CA, MGAZE-Net effectively highlights crucial eye features while considering both local and global spatial relationships. The integration of Global depth-wise Convolution underscores a commitment to nuanced feature extraction. To address the issues of discontinuity and ambiguity in gaze-angle representation, the rotation matrix formalism for gaze ground truth is introduced. This continuous representation ensures a more reliable and consistent estimation of gaze angles. Moreover, the adoption of a 6D rotation matrix representation for direct regression, along with the incorporation of a geodesic-based loss, enhances the balance between accuracy, reliability and efficiency og the gaze estimation.

- **Chapter 6** concludes all proposed methods from previous chapters including L2CS-Net, MTGH-Net, and MGAZE-Net and discuss the possible future research directions of deep gaze estimation in the wild.

2 STATE OF THE ART

This chapter presents a comprehensive overview of state-of-the-art methods in gaze estimation. It explores the transition from conventional eye tracking techniques to the utilization of deep learning for appearance-based gaze estimation. The chapter begins by studying numerous approaches to defining the problem and establishing the particular terminologies used in the field. Next, it focuses on the advances achieved in gaze estimation using deep learning architectures. Finally, the chapter addresses the process of gaze domain adaptation to enhance the performance of gaze estimation using labels from the target domain.

2.1 Introduction

Eye gaze is a pivotal non-verbal communication cue, playing a vital role across a diverse array of applications by providing insights into human intent, cognition [49, 155], and behavior [97, 128]. Its importance extends to the improvement of human-robot interactions [2, 71, 99, 169], the safety of drivers through engagement monitoring [1, 59, 70, 79, 139, 190], and the enrichment of user experiences in virtual and augmented reality environments [39, 150]. These applications depend on accurate gaze estimation methods to interpret the direction and focus of gaze effectively.

The journey of gaze estimation technology is a testament to the remarkable evolution in the field of computer vision. This brief chronology of seminal gaze estimation methods reveals a rich history filled with significant milestones and technological advancements. Figure 2.1 presents a brief chronology of significant innovations in the field of gaze estimation. The foundation for gaze estimation had been laid as early as 1879, using visual observation of the eye's movements while reading a text. Javal et al [90] revealed that reading is made up of saccades, which are quick movements, and fixations, which are short pauses. Shortly thereafter, Delabarre et al [44, 87, 88] developed the first eye movement sensing device using direct mechanical contact with the cornea. However, these methods proved invasive and tended to cause overshooting of eye movements. A key moment in this timeline occurred in 1901 with the introduction of the first non-invasive optical eye tracker by Dodge et al [46]. This technique recorded horizontal eye movement through cornea reflections and required participants to keep their head fixed. This invention laid the groundwork for the field of eye tracking to evolve in various ways. For instance, Buswell et al [18] and Yarbus et al [228] monitored the eye movement of users while looking at complex scenes. They noticed that people mainly focused on meaningful areas for understanding the scene like ob-

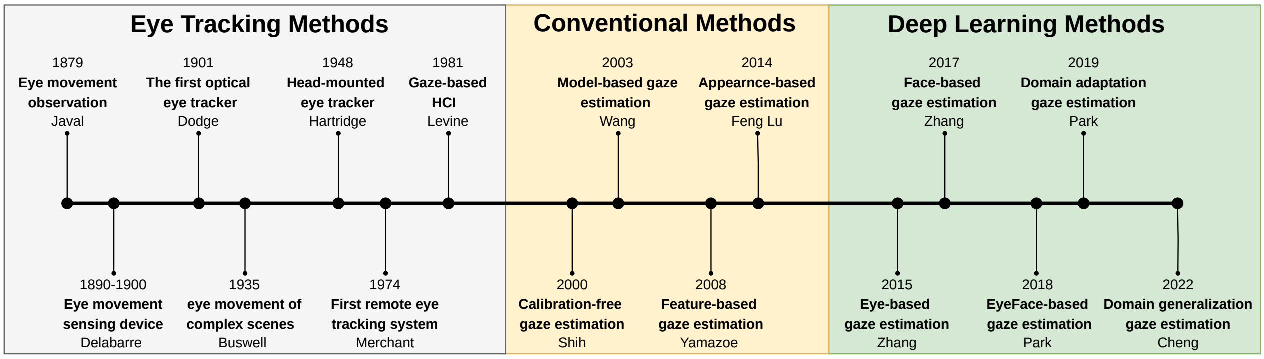

Eye Tracking Methods
Conventional Methods
Deep Learning Methods
1879
Eye movement observation
Javal
1901
The first optical eye tracker
Dodge
1948
Head-mounted eye tracker
Hartridge
1981
Gaze-based HCI
Levine
2003
Model-based gaze estimation
Wang
2014
Appearnce-based gaze estimation
Feng Lu
2017
Face-based gaze estimation
Zhang
2019
Domain adaptation gaze estimation
Park
1890-1900
Eye movement sensing device
Delabarre
1935
eye movement of complex scenes
Buswell
1974
First remote eye tracking system
Merchant
2000
Calibration-free gaze estimation
Shih
2008
Feature-based gaze estimation
Yamazoe
2015
Eye-based gaze estimation
Zhang
2018
EyeFace-based gaze estimation
Park
2022
Domain generalization gaze estimation
Cheng

Figure 2.1: A brief chronology of gaze estimation methods. The first gaze study modelling dates back to the work of Javal et al [90]. The transition from complex hardware to cheaper, off-the-shelf cameras from webcams by Shih et al [165]. One of the first deep learning gaze estimation models was proposed by zhang et al [248].

jects and faces. Hartridge et al [69] introduced the first head-mounted eye tracker that marked a significant step toward allowing greater freedom of head movement in eye tracking studies.

In 1970, gaze estimation and tracking technology flourished. Merchant et al.[129] developed the first eye tracker that remotely measures the direction of the eye without disrupting normal activities. Additionally, Levine et al.[111] utilized newly invented high-speed computers to track gaze in real time, enabling gaze-based human-computer interaction. In the 1990s, Ellis et al.[50] employed eye trackers to examine the usability of websites and emails. However, these early eye tracking devices came with limitations; they required specific hardware and constrained settings to function effectively, which presented a barrier to the widespread adoption of eye trackers as everyday technology for the average user. In the 2000s, the research focus in gaze estimation shifted towards using cheaper, off-the-shelf cameras from webcams or smartphones, thus avoiding the need for specialized devices. Most gaze estimation approaches extract hand-crafted features (e.g. geometric and appearance) from cameras[67, 122, 165, 179, 187, 213, 223, 239] and map these features directly to gaze points or gaze direction.

In 2015, the field of gaze estimation experienced a significant change [54, 107, 248, 250, 252], similar to the transformations seen in other computer vision tasks. This change was driven by the introduction of deep learning techniques. Deep learning, enhanced by access to large amounts of training data, began to overcome issues related to different lighting conditions, camera setups, and the intricate relationship between eye and head movements. Although these advancements significantly enhanced gaze estimation accuracy within consistent datasets, challenges persisted in new environments characterized by variability in facial features, head poses, and lighting conditions. . To address these issues, recent initiatives have focused two main strategies to enhance the adaptability of models across diverse environments and conditions: domain

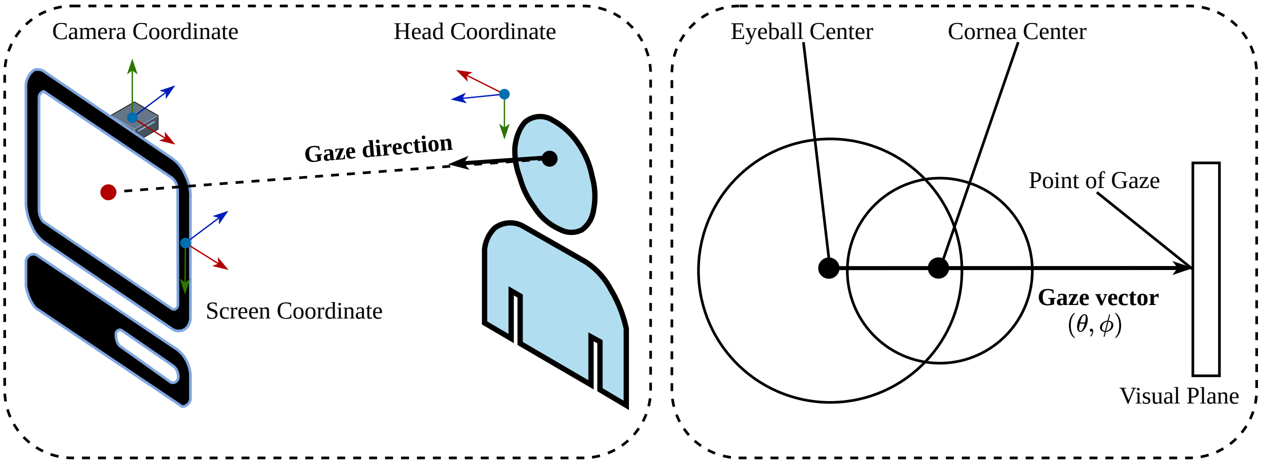


Figure 2.2: In a typical gaze estimation setup, a subject is positioned at a distance from, and directly facing, a camera and screen. The coordinate system is visually organized with color-coded axes: the x-axis in red, the y-axis in green, and the z-axis in blue. The gaze vector is represented by the angles (θ, ϕ) within a polar coordinate system.

generalization [29, 135] and domain adaptation [11, 116].

2.2 Gaze Estimation Problem Setting

The primary objective of gaze estimation is to identify the line of sight, which is the direction in which the pupil is looking. A typical configuration for implementing gaze estimation includes a setup that involves the human subject, sensors, and a visual plane, as illustrated in Figure 2.2. Each component possesses unique characteristics and properties, including:

- **Sensors:** These can be categorized as either intrusive or non-intrusive. Intrusive sensors, while accurate, may cause discomfort as they require physical attachment to the user. Non-intrusive sensors offer a more comfortable experience as they do not need physical contact. However, they encounter challenges such as occlusions, variable lighting conditions, and reflections, which can be particularly problematic in scenarios where the user wears eyeglasses. Over the past decade, a variety of sensors have been investigated in research, including Head Mounted Devices [56, 101, 141, 178], IR Camera [56, 141, 217], Single amera [27, 38, 48, 144, 146, 196, 230, 241, 246–248, 262] and RGB Depth cameras [114], each with its advantages and limitations in different applications.

- **Subject Head Movement:** Accurate gaze estimation extends beyond merely tracking eye movement; it also requires monitoring the orientation of the head. Coordinated movement of the head and eyes in different directions requires gaze tracking systems to consider both factors to ensure accuracy.

- **Visual plane:** The visual plane refers to the surface that includes the point at which the subject's gaze is directed, commonly known as the Point of Gaze (PoG). Typically, gaze estimation systems rely on a camera positioned at a fixed distance from the user. However, in real-world scenarios, the distance between the subject and the visual plane can vary significantly. Common examples of visual planes in gaze estimation setups include computer screens [38, 48, 144, 146, 230, 246, 248] (approximately 60 cm away), mobile phones [59] (around 20 cm away), and Automotive [59, 176] (roughly 50cm away). This variability in the distance to the visual target adds complexity to the design and functionality of gaze estimation systems.

The key calibration factors in such a setup include camera calibration, geometric calibration, and personal calibration. Camera calibration is the essential first step, involving the identification of the camera's intrinsic parameters to ensure the accurate capture of images for analysis. Geometric calibration is the subsequent step, which establishes the precise spatial relationships between the camera, the illumination source, and the display, all crucial for accurately interpreting the captured images. The final step, personal calibration, is conducted to adjust for individual variations, calibrating factors such as head pose and eye-specific attributes, including corneal curvature and the eye's nodal point, to enhance the system's accuracy. For example, CalibMe [162] offers an innovative, fast, and unsupervised technique for calibrating gaze trackers within the realm of human-computer interaction. Accurately predicting the gaze direction involves addressing challenges related to sensor technology, head and eye movement, and variability in the visual target's distance.

With the advancements in gaze estimation systems, a variety of methods have been developed, each offering unique approaches and applications. Numerous studies [5, 34, 57, 103, 148] have explored eye gaze estimation research, attempting to classify the diverse methodologies employed. This task is complex due to the often ambiguous distinctions between methods and the common use of hybrid techniques. For instance, Kar et al.[93] categorize these methods into five types, including 2D regression, 3D modeling, appearance-based, cross-ratio-based, and shape-based methods. Another categorization, proposed by Cazzato et al.[23], differentiates between geometric-based and appearance-based methods. This chapter synthesizes insights from these reviews to categorize gaze estimation methods into two principal categories: conventional and deep learning methods.

2.3 Conventional gaze estimation

Conventional gaze estimation use images directly from cameras and map them to the gaze direction. This way, gaze estimation could be used easily in unconstrained settings without complex hardware. In this section, three conventional gaze estimation

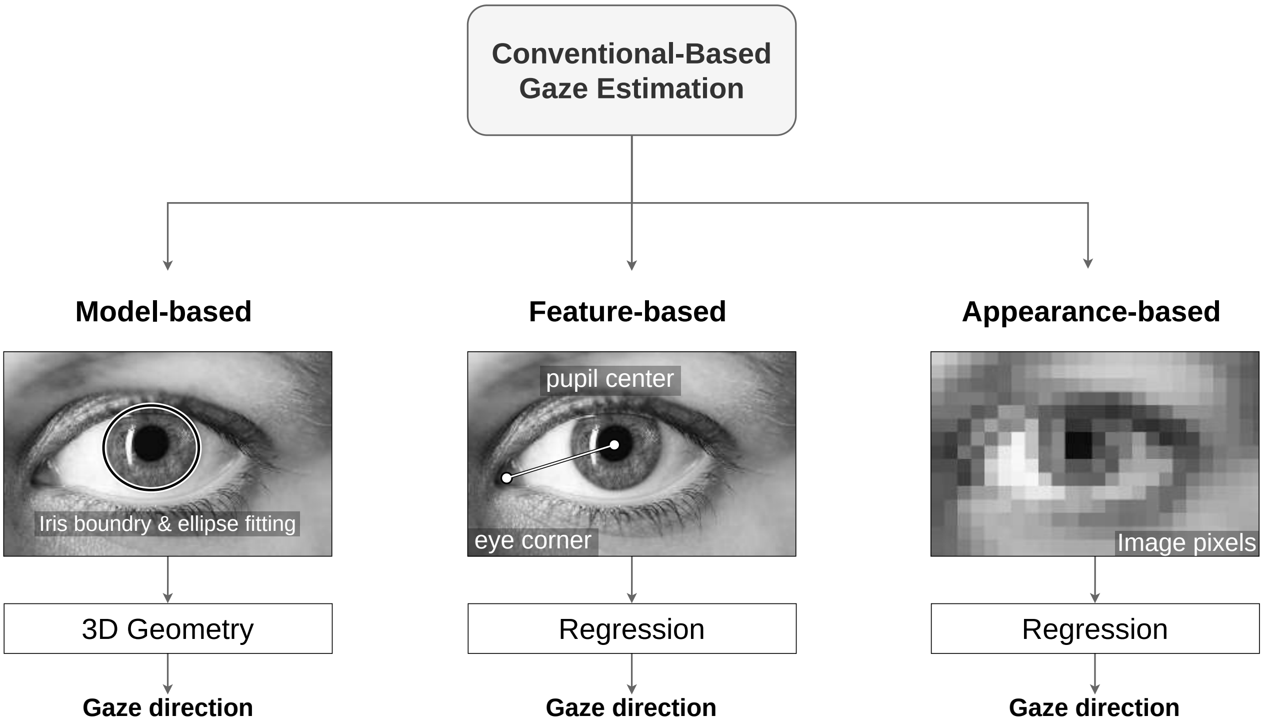


Figure 2.3: Conventional-based gaze estimation methods include three main categories. model-based methods rely on 3D geometry, feature-based and appearance-based methods rely on image pixels.

categories are presented as shown in Figure 2.3. These categories include model-based, feature-based and appearance-based. The recent developments are exploreed within each of these categories in the following subsections.

2.3.1 Model-based Gaze Estimation

Model-based methods in gaze estimation utilize the anatomy of the eyeball through two types of models include elliptical iris boundary and spherical eyeball models. Elliptical iris boundary models operate on the principle that the iris, which is circular in three dimensions, appears as an ellipse in two-dimensional images. This process involves two main steps: initially capturing the 2D limbal ellipse from the eye image, and then projecting it back into its original 3D circular form. The gaze direction is determined using the normal vector of the 3D circle. To accurately project the 2D ellipse of the iris into a 3D circle, the radius of the iris must be known. For instance, Wang et al.[188] utilized the Canny algorithm[20] to accurately identify the edges of the iris using a high-resolution image captured with a zoom-in camera. They fit the iris contour to an ellipse, then projected it to the 3D circle, representing iris radius as average anatomical values. Hansen et al.[68] developed an active iris tracker using particle filtering that demonstrated robustness against various changes in lighting and camera settings. Zang et al.[239] proposed an improved and reliable RANSAC algorithm to enhance the accuracy of limbal ellipse extraction. Tsukada et al.[179] determined gaze direction by accurately fitting the iris shape and employing a lookup table that correlates different iris shapes with their respective gaze directions. Wood et al.[213] estimate gaze direction using a regular tablet with a built-in front-facing camera, based solely on eye geometry. They identified a set of possible limbus edge points to avoid the computational costs associated with popular iris boundary detectors. Despite these advancements, challenges such as partial eyelid closure still pose significant hurdles in accurately extracting the limbal ellipse from images.

In contrast, spherical eyeball models identify the position of the pupil and the center of the eyeball in 3D space, using the line connecting these points to predict gaze direction. Various techniques have been employed to determine the center of the iris and the eyeball. Valent et al.[183] and Timm et al.[183] employed shape-based methods to estimate the center of the iris. Yamazoe et al.[223] created a 3D model of a face over time by tracking key points in example images to estimate the position of the head and the center of the iris. Xiong et al.[218] proposed a face model for each individual, manually setting the position of the eyeballs, and then mapping the coordinates of the detected iris center onto this pre-established 3D model. On the other hand, the position of the 3D eyeball center was estimated using head tracking [6], facial landmarks [8, 89], or eyelid contours [214]. Furthermore, recent studies have explored the fitting of morphable models directly to the eye region, optimizing appearance and illumination [195, 210], advancements that aid not only in estimating gaze direction but also in gaze redirection [212]. The use of additional light sources [62, 165] and multiple cameras [229] has been explored to improve robustness against head movements and to provide more accurate reflections of the corneal surface. Despite these advances, uncertainties remain in determining the 3D center of the eyeball, affecting the accuracy of gaze estimation.

2.3.2 Feature-based Gaze Estimation

Feature-based gaze estimation techniques focus on identifying and detecting specific key points within the input image of the eye. These methods directly utilize identified features in regression-based techniques, offering robustness against challenges like changes in lighting or the subject's head pose. These techniques typically depend on the location of key points in the eye image, such as the offset between the pupil center and corneal reflection (PC-CR), which necessitates the user's head to be fixed and directly facing the computer screen. Cerrolaza et al.[24] demonstrated that a single infrared source is sufficient to track stationary gaze using the PC-CR vector, and two IR sources are required for tracking gaze with dynamic head movements. Another essential feature utilized is the pupil center and eye corner (PC-EC) vector, which measures the distance between the center of the pupil and the inner corner of the eye. Sesma et al.[163] showed that the PC-EC vector could effectively replace the PC-CR vector in webcam settings. This approach is practical for implementation, as evidenced by its widespread use in many webcam-based eye-tracking systems [83, 254, 255], since it does not require meticulously calibrated hardware. Specifically, Hansen et al.[66] employed an active appearance model to identify the PC-EC vector and applied a gaussian process to estimate gaze direction. Valenti et al.[184] adapted the isophote curvature technique to detect PC-EC locations and used a linear mapping method to predict gaze.

To enhance reliability across various gaze values and head poses, researchers have investigated the inclusion of a more extensive set of feature points in the image. For instance, Bäck et al.[7] augmented the traditional PC-EC features with additional data points, such as nostril positions and head pose, creating a more comprehensive feature vector for regression. Similarly, Skodras et al.[167] enhanced their gaze tracking methodology by incorporating points on the eyelids, which significantly improved the accuracy of vertical gaze tracking. Huang et al [84] learned a gaze model from numerous daily human-computer interactions by adapting additional gaze points, eye movement patterns, and head pose data. Despite the robustness to certain environmental and individual variations, feature-based gaze estimation still faces challenges, particularly in accurately localizing features under diverse conditions and in the presence of occlusions or extreme head movements. The performance of these systems heavily depends on the selection and reliability of the feature points.

2.3.3 Appearance-based Gaze Estimation

The field of gaze estimation has experienced a significant transformation with the introduction of appearance-based methods. Traditional model-based or feature-based techniques rely on specific feature points or facial landmarks within the eye image. For instance, model-based methods depend on tracking the limbus, but these struggle when the eyes are nearly closed or partially obscured. Similarly, feature-based methods often focus on tracking the corners of the eyes, facing challenges with the inner eye corner a poorly defined landmark that is difficult to accurately track. In contrast, appearance-based approaches offer a solution by conducting a holistic analysis of the entire eye image, addressing the issues associated with specific feature tracking. This method transforms eye image pixels into a feature space, which is then mapped to gaze direction. This shift emphasizes a data-driven learning process rather than reliance on predefined eye features. Early implementations of this methodology include the work of Baluja and Pomerleau [9, 151], who trained a neural network with 2,000 training samples to predict the position of gaze on a screen. Later, Tan et al [173] developed a method based on pixel similarity, using a nearest-neighbor approach with 252 labeled samples to predict screen coordinates.

Early appearance-based methods often focused on learning a mapping function specific to each subject, requiring a time-consuming calibration process to gather the necessary training samples for that individual. To minimize the need for extensive training samples, Williams et al.[208] introduced the use of semi-supervised gaussian process regression methods. Additionally, Sugano et al.[170] developed a novel approach by integrating gaze estimation with saliency models, offering a more efficient and potentially more accurate method to understand gaze behavior. Lu et al [121] reduced the resolution of eye images to $3x5$ pixels and applied adaptive linear regression to identify the most effective set of sparse training samples for interpolation. While these methods demonstrated acceptable performance within constrained settings (head pose is fixed and the subjects are consistent), their effectiveness significantly diminishes in unconstrained environments. To address the issue of performance decrease across different subjects, Funes et al.[131] introduced a method of cross-subject training, though the average error reported by this approach exceeds 10 degrees. Sugano et al.[171] proposed a learning-by-synthesis strategy, utilizing a vast collection of synthetic cross-subject data to train their model. Furthermore, Lu et al [118] adopted a sparse auto encoder approach, learning a set of basis functions from patches of eye images and using these bases to reconstruct the eye images. To manage the challenge of head movement, Sugano et al.[172] grouped training samples that share similar head poses and performed gaze interpolation within a local manifold. Meanwhile, Lu et al.[120] recommended starting the estimation process with the original training images and adjusting for any bias through regression. Additionally, Lu et al [123] introduced an innovative method for gaze estimation that accommodates free head movement by synthesizing eye images with a single camera.

Another direction in appearance-based gaze estimation is the exploration of more sophisticated feature spaces. Schneider et al.[119] investigated various features, including the discrete cosine transform (DCT), local binary patterns (LBP), and histogram of oriented gradients (HOG), employing dimensionality reduction methods to enhance accuracy. Huang et al.[85] found that using multilevel HOG features combined with a random forest yielded the most effective accuracy. However, these methods typically perform well in controlled settings with a specific subject or fixed head pose but see a decline in accuracy in unconstrained environments. A significant advancement in appearance-based gaze estimation was achieved through the acquisition of large training data collected with a multi-view camera system, coupled with 3D reconstruction techniques to create a synthetic dataset via structure-from-motion [171]. This approach utilized random forest regression on the synthesized large-scale dataset to achieve accurate cross-person gaze estimation without the need for individual calibration steps. Moreover, the advent of appearance-based methods has significantly improved the applicability and consistency of experimental results in gaze estimation research. By harnessing the power of large-scale data and leveraging modern machine learning capabilities, this approach has opened new avenues for developing more accurate, robust, and user-friendly gaze tracking technologies.

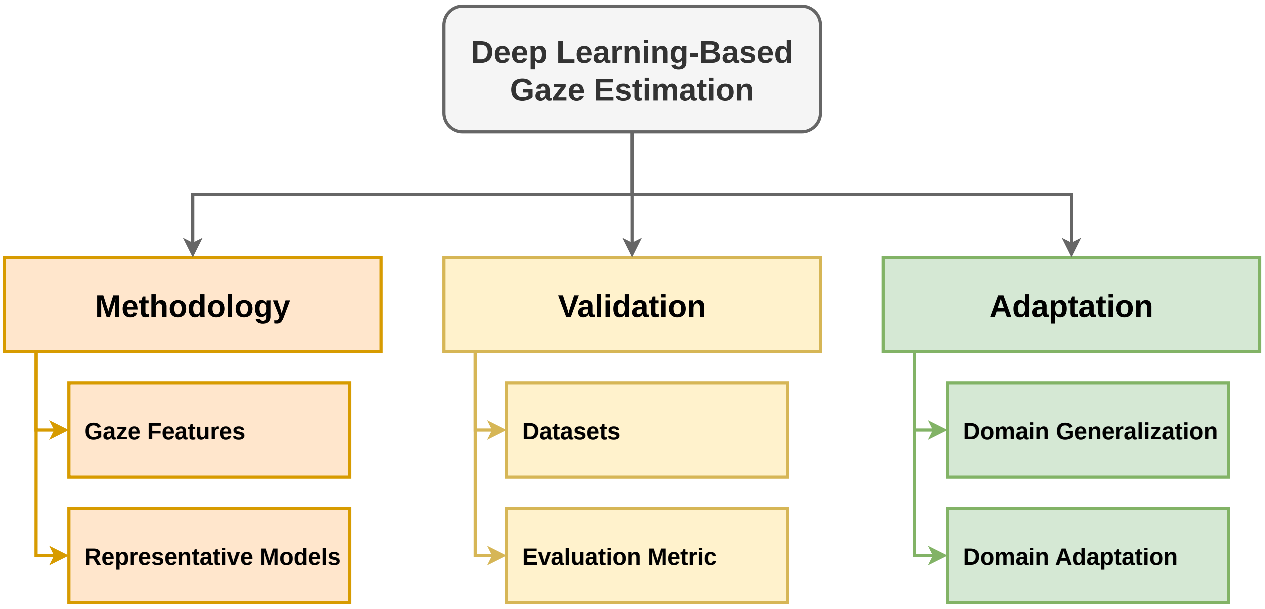


Figure 2.4: Deep learning gaze estimation survey from three perspectives including methodology, validation, and adaptation.

2.3.4 Summary

Conventional methods for gaze estimation fall primarily into three categories: model-based, feature-based, and appearance-based approaches. Model-based methods rely on geometric and anatomical models of the eye, such as elliptical iris boundary and spherical eyeball models, to estimate gaze direction from images. Feature-based methods identify specific features or key points in the eye image, such as the pupil center or eye corners, and use these for gaze estimation through regression techniques, showing robustness to variations in lighting and head pose. Appearance-based methods, on the other hand, analyze the whole eye image, employing machine learning techniques to directly map the eye appearance to gaze direction, overcoming the limitations of tracking specific features. These methods have evolved to handle challenges like occlusions, head movements, and variability across different individuals, with appearance-based approaches significantly benefiting from large-scale data and advanced computational models to achieve more accurate and user-friendly gaze tracking.

2.4 Deep learning Gaze Estimation

Appearance-based gaze estimation faces significant challenges, including variations due to individual differences and head movements in unconstrained settings. Fortunately, the advent of deep learning, particularly Convolutional Neural Networks (CNNs), has offered a robust solution by leveraging their superior performance across various computer vision tasks to effectively address these challenges. The field has seen an increase in publicly available datasets and innovative methods that integrate insights from broader areas of deep learning and CNNs. Among these methods, Zhang et al [248] introduced the first CNN-based model for appearance-based gaze estimation. They trained a deep CNN with 200,000 images from 15 individuals captured during daily laptop use. Their work significantly improved upon previous methods in gaze estimation, paving the way for deeper exploration of Deep learning techniques in this area. This section surveys current Deep learning gaze estimation methods from three perspectives: methodology, validation, and adaptation, as depicted in Figure 2.4. This framework provides a structured overview of the state-of-the-art, facilitating further research and development in the field.

2.4.1 Methodology

Deep learning gaze estimation methodology involves two major steps: feature extraction and model development. The first step entails choosing the type of input images to be used, which can include eye images, face images, or both. This step focuses on extracting crucial gaze features from these images, taking into account the complex appearance of the eye and other relevant facial features. The second step involves proposing models that accurately derive gaze information from the extracted features. These models transform the feature space from the images into precise gaze predictions. The accuracy of these models largely depends on the type of input used and the architectural design of the models employed to process this input.

2.4.1.1 Gaze Features

Gaze feature extraction are categorized based on the types of input images into three groups: eye images only, eye and face images, and face images only, as illustrated in Figure 2.5. This classification aids in understanding how different inputs can be utilized to enhance the performance of gaze estimation models by leveraging their respective strengths in representing gaze-relevant information.

1) Features from solely eye images: the relationship between gaze direction and eye appearance is fundamental to gaze estimation. Changes in gaze direction result in observable modifications in the eye, such as shifts in the location of the iris and changes in the shape of the eyelid. Consequently, conventional gaze estimation methods [124, 173] have relied on features extracted directly from eye images. However, these features often exhibit high redundancy and face challenges in adapting to changes in the environment. Deep learning methods have significantly advanced the field of gaze estimation by automatically extracting gaze features from eye images. Initially, Zhang et al.[248] extracted features from single-eye gray scale images using a LeNet-based approach, which utilizes a CNN to enhance feature extraction by incorporating estimated head pose information. Park et al.[146] transformed the image of the single eye into a unified gaze representation, encapsulating detailed pictorial depictions of the

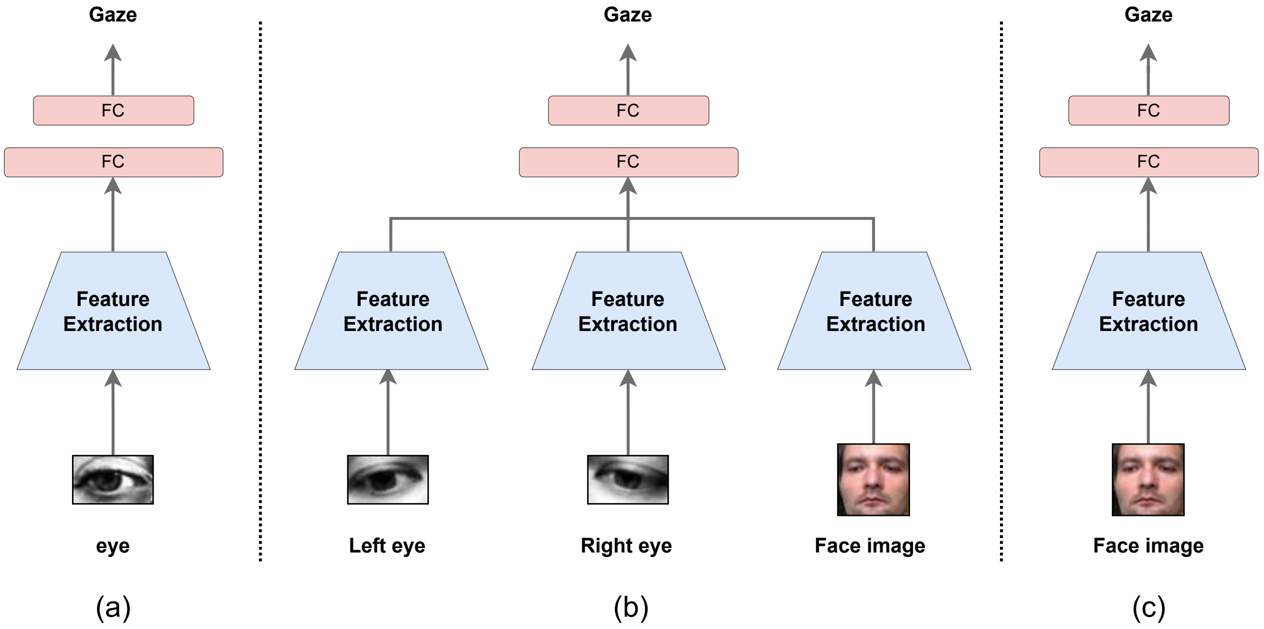


Figure 2.5: Gaze estimation categories based on the types of input image include. (a) Features from solely eye images [82, 146, 248, 252]; (b) Features from combination of eye and face images [27, 30, 132, 215]; and (c) Features from solely face images [29, 135, 194, 250]

eyeball, iris, and pupil. They employed a lightweight DenseNet[82] to regress the gaze direction from the pictorial representation, facilitating a more intuitive regression of gaze direction directly from these representations, thereby enhancing the accuracy and interpretability of gaze estimation. The evolution of gaze estimation techniques has also seen a shift from analyzing single-eye images to incorporating features from both eyes to enhance accuracy. Cheng et al [33] developed a novel four-stream CNN network for eye gaze estimation, integrating with the Asymmetric Regression-Evaluation Network (ARE-Net). This architecture utilizes two branches to extract features from the left and right eyes independently, while the other two branches are dedicated to analyzing joint features.

Huang et al [86] enhanced mobile gaze estimation by leveraging both eyes and implementing a baseline algorithm, achieving satisfactory accuracy on a tablet screen. They further evaluated the performance of four different regression algorithms, including k-Nearest Neighbors (kNN), Random Forests (RF), Gaussian Process Regression (GPR), and Support Vector Regression (SVR), providing invaluable insights for future advancements in gaze estimation accuracy. Fischer et al.[54, 147] extract individual features from each eye image using two networks utilizing VGG-16 architecture. Generative Adversarial Networks (GANs) are employed to preprocess eye images, addressing environmental challenges specific to gaze estimation. Kim et al.[102] enhanced eye images captured under poor lighting conditions using the EnlightenGAN framework[91], which helps recover vital information lost in low light, thus improving gaze estimation performance. Additionally, Rangesh et al [152] utilized a GAN to remove eyeglasses from eye images, significantly reducing potential reflections and distortions. Gaze direction is influenced by both the head's position and orientation, as well as the movement of the eyeball and pupil. Modern gaze estimation methods, increasingly, consider both eye gaze and head pose simultaneously for more accurate assessments [54, 248]. However, integrating head pose data into gaze estimation algorithms remains challenging, often necessitating the use of existing head pose estimation systems or manually-engineered features to ensure precise predictions.

2) Features from combination of eye and face images: to reduce the complexity of incorporating head pose information in gaze estimation, several approaches combine eye features with the entire facial image as inputs to their models, which naturally contains crucial head pose information. For example, Karfa et al [107] introduced the iTracker, an eye-tracking system based on a CNN that processes multiple inputs to estimate gaze direction. This system utilizes images of the left eye, right eye, and face, all cropped from the original frame. Additionally, it incorporates a face grid input, which is a binary mask that indicates the location and size of the head within the frame. This multi-input approach allows iTracker to accurately determine gaze direction by analyzing both the detailed features of the eyes and face, as well as the position of the head in the frame. Chen et al.[27] extracted features from images of the face and both eyes separately, employing dedicated networks for the face and each eye. They utilized dilated convolutions to extract high-resolution features, effectively enlarging the receptive field of convolutional filters without compromising spatial resolution. In subsequent work, Chen et al.[30] estimated a basic gaze direction from the face image, which contains richer information, and refine this estimate using fine-grained features extracted from eye images. They also proposed FARE-Net [36], which uses a face-based asymmetric regression network to predict 3D gaze directions from face and eye images, complemented by an evaluation network that adaptively optimizes this process by assessing the reliability of each eye.

Murthy et al.[132] integrated features extracted from both eyes and the entire face into the final prediction layers. Unlike traditional approaches that merge face and eye features, Wu et al.[215] improve gaze estimation by using eye features to refine face features. Zhou et al [261] initially extract facial features to guide the extraction of left and right eye features separately, then iteratively compute and correlate contextual features for both eyes using Correlation-based Blocks (CCBs), which adaptively assign shared attention weights. Murthy et al.[133] extracted feature maps for images of left and right eyes using shared convolutional layers. They captured high-level information such as the eyeball, sclera, and brow regions, while retaining the face channel essential for encoding head pose information. Facial landmarks also enhance gaze estimation by modeling head pose and eye position. Palmero et al.[140] integrated face, eye regions, and facial landmarks in a CNN to improve accuracy. Dias et al.[45]

used facial landmarks to directly predict gaze direction and incorporate prediction uncertainty into their output. Jyoti et al.[92] extracted geometric features, such as angles from the pupil center to key facial landmarks, to refine gaze estimation. Furthermore, Dubey et al [48] leveraged detected facial landmarks for unsupervised gaze representation learning, classifying gaze zones in web-collected images, showcasing the utility of landmarks in various gaze estimation tasks. Although integrating facial images with eye images enhances the accuracy and robustness of gaze estimation, it adds computational complexity to the network's training and inference processes.

3) Features from solely face images: using eye images in gaze estimation typically involves two major stages: initially segmenting the eye area from the face and then deducing the gaze direction from that area. However, this approach encounters several challenges. Eye segmentation requires additional annotations such as eye location or head orientation, alongside gaze ground truth [54, 107, 252]. Inaccuracies in isolating the eye area can compromise gaze deduction. Furthermore, performing segmentation at the inference stage adds complexity and does not ensure an optimal solution for accurate gaze estimation. Recent studies have adopted an approach where solely face images are directly used as input for gaze estimation models [250]. Using the entire face image, which contains critical gaze information from the eyeballs and head pose, is vital for accurate gaze estimation. Moreover, utilizing the entire face image as the sole input simplifies the architecture of the gaze estimation model, reducing its complexity. This approach enhances the effectiveness and efficiency of gaze estimation by leveraging the natural synergy between eyeball orientation and head pose in predicting gaze direction. Zhang et al.[250] introduced the first full-face CNN architecture for gaze estimation that differs from traditional methods by using the entire face image to directly predict gaze direction. They outperformed existing eye-only and multi-input methods, achieving significant improvements in gaze estimation accuracy.

Researchers have developed methods to filter out redundant information from face images to enhance the efficiency and accuracy of gaze estimation [136, 250]. Zhang et al [250] integrated a spatial weighting mechanism into standard CNN architectures to prioritize significant facial regions by learning spatial weights from the activation maps of the convolutional layer, effectively reducing noise. Zhang et al [247] proposed a learning-based approach for dynamically selecting optimal facial sub-regions for gaze estimation. Cheng et al[29] developed a self-adversarial network designed as a plug-and-play solution to enhance facial features by eliminating irrelevant details while preserving crucial gaze-related information. Oh et al [135] introduced a novel projection that applies convolution across the full face image to accurately model local context while reducing the computational cost of self-attention. The proposed model incorporates deconvolution to upscale the down-sampled global context back to the input size to preserve spatial information. To address the challenge of estimating gaze from low-resolution facial images, Zhu et al[270] introduced a novel approach that leverages the consistent structure of faces to reconstruct missing information through an end-to-end mapping from low to high-resolution images. Wang et al[194] proposed the GazeCaps framework, which uses capsules to encode diverse facial properties for gaze estimation. These capsules adeptly adjust to facial transformations through vector expressions, effectively managing the nonlinear changes in facial components influenced by head direction and perspective.

Although full-face gaze estimation addresses some issues found in eye-based and multi-input methods, it also encounters its own set of challenges, including the fact that the eyeball, which is critical for gaze information, occupies only a small portion of the face image. Traditional CNNs often struggle to extract fine-grained gaze features without also capturing irrelevant facial features. These challenges pave the way for the development of advanced deep learning methods aimed at improving face-based gaze estimation.

2.4.1.2 Representative models

The function of representative models in gaze estimation is to extract crucial gaze features from images and map them to gaze directions. Within the realm of deep learning, these models are categorized into five categories: CNN-based, temporal-based, generative-based, transformer-based, and lightweight models.

1) CNN-based gaze estimation models: CNNs have emerged as a powerful tool in computer vision, delivering impressive results across a range of tasks, including gaze estimation [3, 130, 138, 226, 227, 238]. For gaze estimation, CNNs are categorized into two main classes: supervised CNNs and semi/self/unsupervised CNNs. The typical architecture of a supervised gaze estimation CNN is designed to train with image samples annotated with ground-truth gaze values. Early approaches in gaze estimation leverage the supervised training paradigm, focusing on the CNN architecture to extract essential gaze information. These methods utilize conventional CNN architectures with well-established backbones such as LeNet [248], Alex-Net [107], VGG [54], ResNet-18 [95], and ResNet-50 [242]. Alternatively, some strategies involve specially designed CNN architectures [30, 36, 113, 230] that are tailored to precisely extract fine-grained gaze features from images.

GazeNet [248], utilizes a LeNet-based architecture, comprises five convolutional layers followed by two FC. This model processes gray scale eye patch images to map them onto a 2D angular gaze vector. An improved version of this model was presented in a subsequent study [252], employs a deeper architecture with 13 convolutional layers to achieve improved performance.

FullFaceNet [251], employs AlexNet backbone, consisting of five convolutional layers and two FC as shown in Figure 2.6. This model enhances its capabilities by

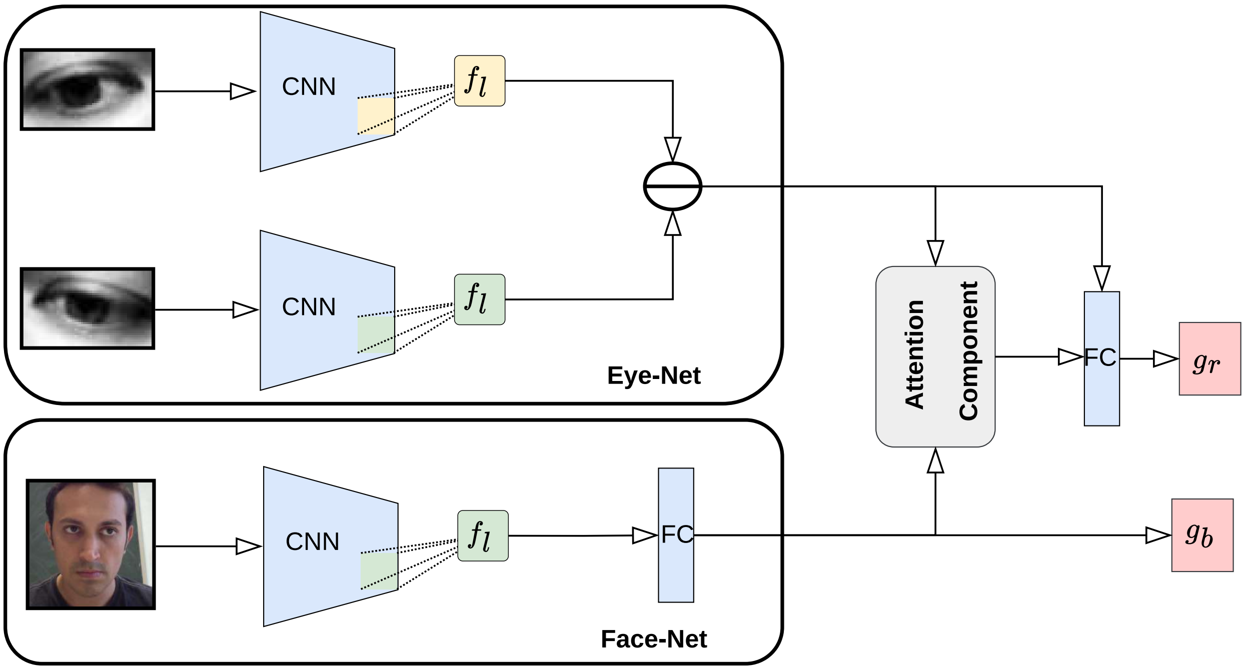


Figure 2.6: CA-Net, a supervised CNN model developed by Cheng et al [30] utilizes two specialized sub-networks: Face-Net and Eye-Net. Face-Net employs a CNN composed of 13 convolutional blocks to process face images and estimate basic gaze directions. Meanwhile, Eye-Net focuses on the left eye and features a CNN with 10 convolutional blocks, enhancing the model's capability to analyze and interpret detailed gaze data effectively.

incorporating a spatial weighting mechanism, which includes three additional 1 × 1 convolutional layers followed by ReLU activation. This model integrates a spatial weighting mechanism by adding three 1 × 1 convolutional layers followed by ReLU activation. The purpose of this enhancement is to emphasize crucial areas within the facial images by adjusting the weights of the activation maps.
RT-Gene [54], uses VGG-16 network to process eye patches and extract gaze features. Following this feature extraction phase, the output from each VGG-16 network is directed into a FC layer, which is then subjected to batch normalization and ReLU activation. The head pose vector is incorporated at this point, and this is followed by two additional FC layers that yield the gaze estimation.

CA-Net [30], adapts a coarse-to-fine strategy equipped with an attention module to capture fine-grained gaze features. This approach utilizes two sub networks: Face-Net and Eye-Net. Face-Net employs a CNN with 13 convolutional blocks to process face images and estimate basic gaze directions. Meanwhile, Eye-Net, focusing on the left eye, is equipped with a CNN featuring 10 convolutional blocks, enhancing the model's ability to analyze and interpret gaze data effectively.
FARE-Net [36], features two subnets: FARNet and ENet. FARNet, a three-stream CNN, processes images from the left and right eyes with identical six-layer architec-

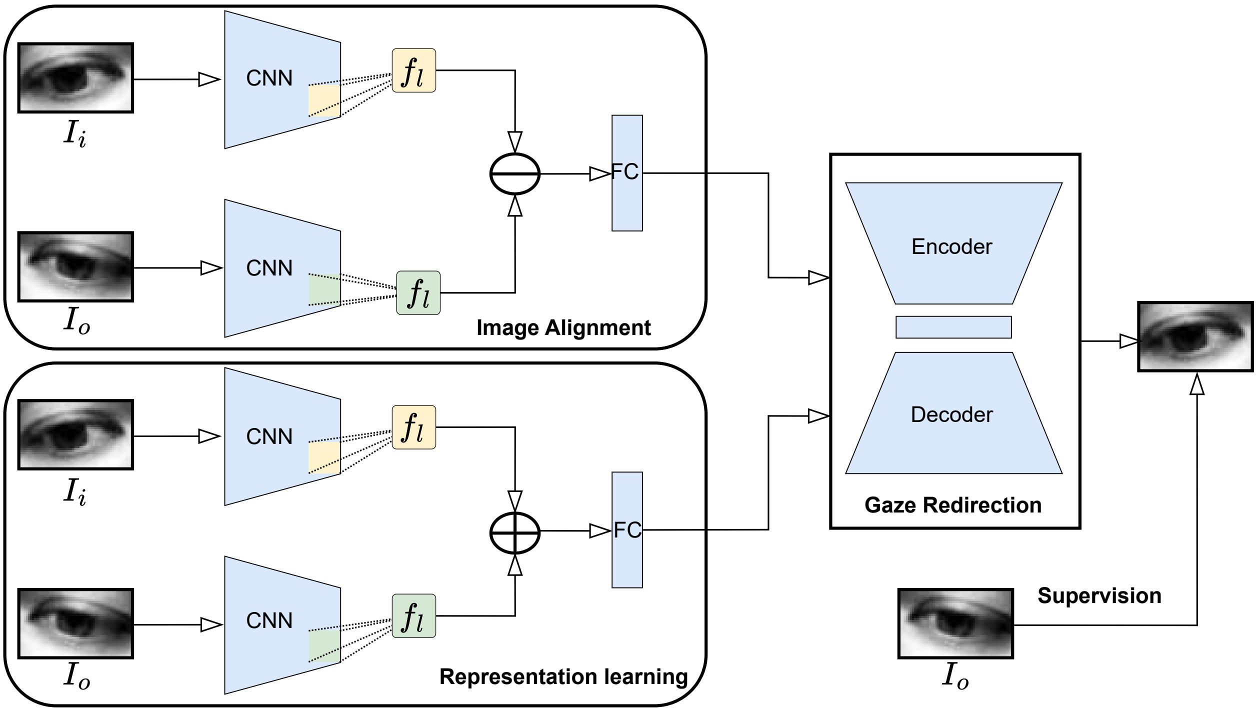


Figure 2.7: The unsupervised CNN model developed by Yu et al [232] utilizes a gaze redirection network and leverages the discrepancy in gaze representation between input and target images. This framework facilitates the concurrent training of both the redirection and representation networks by employing a redirection loss in the image domain, eliminating the need for gaze annotations.

tures, followed by a stream utilizing AlexNet for facial image processing. ENet, a two-stream CNN, processes left and right eye images separately, each through six convolutional layers and two fully connected layers.

MTMV-Net [113], estimates gaze directions from eye images and the point of gaze (PoG) from both single and multi-view eye images. This network also includes a task for PoG estimation using depth images integrated with facial features, demonstrating a comprehensive approach for accurate gaze detection.

LGM-Net [230], combines eye landmark locations with gaze directions in a multi-task CNN framework. This model effectively decodes alignment parameters for eye landmarks to calculate gaze direction, providing an integrated gaze estimation solution. These various architectures demonstrate the dynamic evolution of supervised CNN-based gaze estimation, reflecting continuous improvements in accuracy, efficiency, and application adaptability. The accuracy of supervised deep learning in gaze estimation is significantly dependent on the availability of a large scale of annotated data. However, obtaining accurate gaze annotations is both complex and time-consuming. To address this challenge, Semi/Self/Un Supervised CNNs emerge as promising solutions. These approaches utilize unlabeled images to improve performance, taking advantage of the lower cost and ease of gathering such data.

Semi-supervised CNNs, utilize a combination of labeled and unlabeled images to optimize network training. Wang et al [197] introduced an adversarial learning approach for semi-supervised learning, specifically designed to enhance model performance for particular subjects or datasets. This method combines the strengths of both supervised and unsupervised learning to improve generalization across varied data conditions.

Self-supervised CNNs, employ pretext auxiliary tasks to enhance gaze estimation performance without direct gaze annotations. Cheng et al [33] introduced a self-supervised asymmetry regression network, featuring a regression component tasked with determining the gaze directions of both eyes. This model also includes an evaluation mechanism that assesses the reliability of gaze estimations for each eye, ensuring both accurate and consistent predictions.

Weakly supervised CNNs, combine elements of both supervised and unsupervised methodologies. They primarily use limited labeled data to augment the learning process with predominantly unlabeled data. Kothari et al [60] leveraged the LAEO (looking at each other) dataset[106], utilizing geometric constraints related to gaze in two-person interactions to introduce a unique perspective on weak supervision. Similarly, Ghosh et al [105] propose the MTGLS framework, which exploits non-annotated facial images by integrating three signals: the pupil's line of sight, head pose, and eye movement, to enhance the training process.

Unsupervised CNNs, which are trained using entirely unlabeled data, face significant challenges due to the absence of ground-truth data. As shown in Figure 2.7, Dubey et al [48] address this by collecting unlabeled facial images from web pages and approximating gaze region annotations based on detected facial landmarks. This innovative approach allows them to undertake gaze representation learning as if it were a supervised task, making strides towards effective unsupervised gaze estimation.

2) Temporal-based gaze estimation models: The human gaze is inherently dynamic, exhibiting complex movements with a strong temporal correlation between successive frames. Recognizing this, numerous studies [95, 140, 142, 196, 262] have used temporal information to significantly improve the accuracy of gaze estimation, surpassing methods that rely solely on static images. Various temporal architectures, including GRUs [140], LSTMs [262], and bi-LSTMs [95], have played a crucial role in these developments.

iTracker-improved [262], uses an innovative enhancement to the iTracker framework by incorporating a Bi-LSTM network to effectively harness temporal sequence information from consecutive frames which enhance gaze estimation accuracy. This approach effectively combines static image analysis with dynamic temporal data, offering a comprehensive and advanced solution for gaze estimation.

MT-Modal-Recurrent [262], utilizes a multimodal recurrent CNN framework that processes invariant features from all input frames in a sequence, subsequently inte-

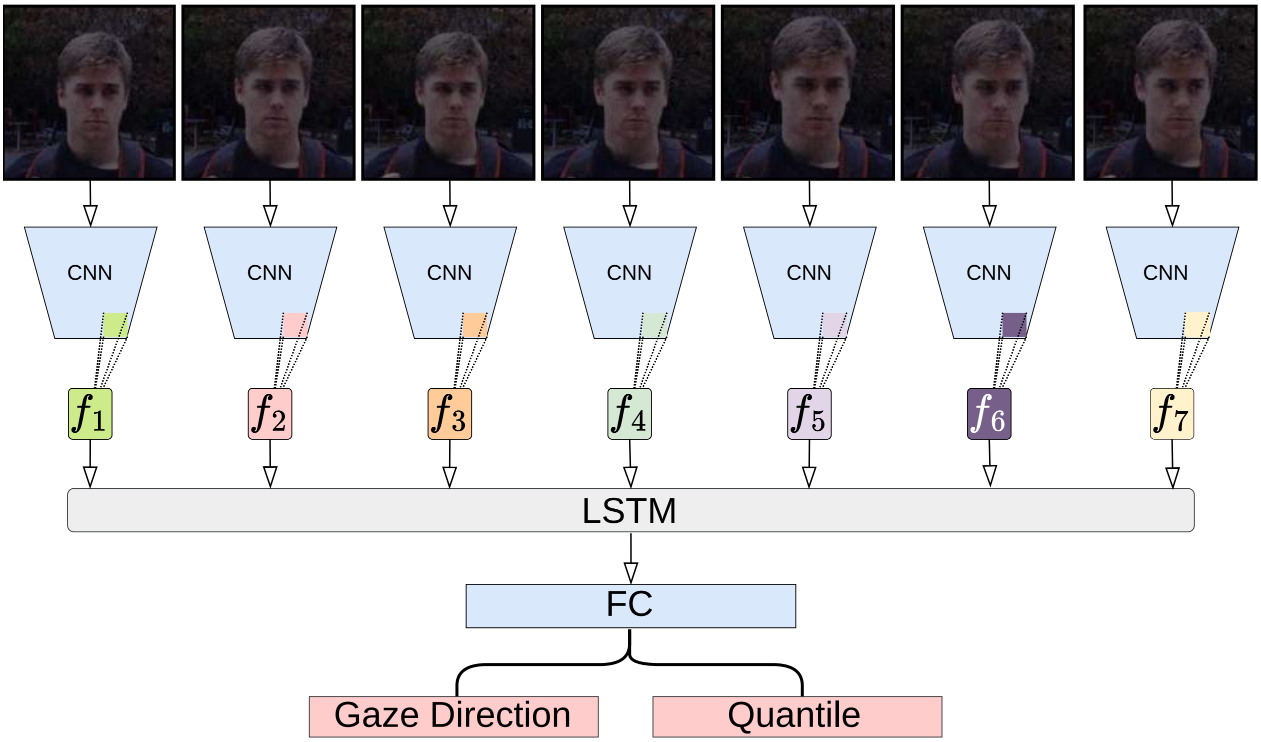


Figure 2.8: The temporal model developed by Kellnhofer et al [95] combines CNNs and bi-LSTM networks to predict gaze direction from video sequences. It analyzes seven frames at a time, using CNNs for feature extraction and bi-LSTMs to incorporate temporal dynamics, thereby enabling accurate gaze predictions.

grating these through a dedicated multimodal recurrent module to predict the 3D orientation of gaze in the final frame. While their findings indicated no significant improvements for on-screen target gaze estimation, the results showed promise for smooth pursuit of floating targets within the EYEDIAP dataset, suggesting that temporal data may not substantially enhance gaze estimation in simpler scenarios such as screen tracking but could be beneficial in more complex situations involving varied gaze directions and head movements.

Gaze360-LSTM [95], integrates CNNs and bidirectional Long Short-Term Memory bi-LSTM) network for the tracking and prediction of gaze direction from video sequences. By analyzing a sequence of seven frames, the model utilizes a CNN to extract high-level features from each frame and then employ bi-LSTM to process these features over time as shown in Figure 2.8. The bi-LSTM architecture is key to leveraging both past and future contextual information to enhance prediction accuracy. Additionally, the model output includes an estimate of the prediction error, which demonstrates its ability to understand and predict human gaze behavior accurately.

Spatio-Temporal [142], employs a spatio-temporal model that compares a recurrent CNN approach with a static only CNN model, underscoring the critical role of temporal data in the accurate tracking of eye movements. The integration of temporal information from high frame rate eye image sequences significantly improves the ac-

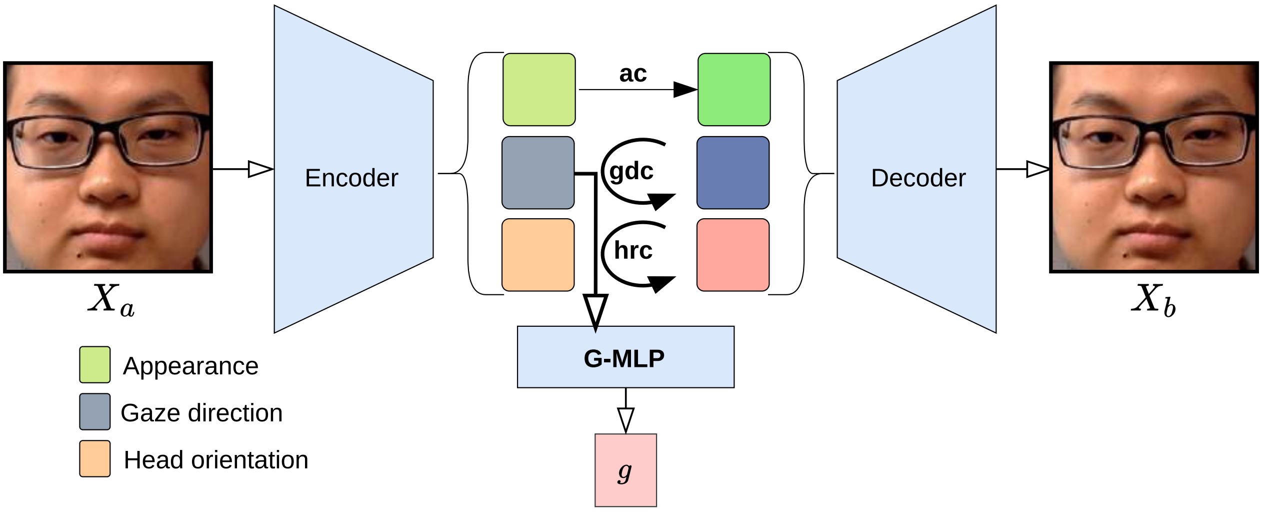


Figure 2.9: The generative model developed by Park et al [144], named FAZE, uses a DT-ED to extract key gaze-related features such as gaze direction, head pose, and eye appearance into a structured latent space.

curacy of gaze estimation, especially in tracking the vertical component of gaze. This finding is particularly relevant for applications involving virtual reality head-mounted displays. Although these studies [95, 140, 142, 196, 262] confirm the benefits of temporal information in gaze estimation, challenges remain in capturing eye dynamics in real-world scenarios and low-quality videos. More research is needed to understand eye movement patterns across different tasks and natural behaviors, highlighting significant gaps in current understanding.

3) Generative-based gaze estimation models: The exploration of GANs has opened new avenues in the field of gaze estimation. These models have shown significant promise in learning complex, high-dimensional data distributions, which is essential for accurately inferring gaze direction from visual inputs.
DT-ED framework [144], leverages the Disentangling Transforming Encoder-Decoder (DT-ED) to learn latent representations of crucial gaze-related features from input images, such as gaze direction, head pose, and eye appearance. This framework maps these inputs to a well-structured latent space to facilitate accurate and efficient gaze estimation as shown in Figure 2.9. This approach increases the model's generalizability across various subjects by facilitating a detailed understanding and manipulation of gaze that is independent of person-specific attributes.
ST-ED framework [258], leverages the Self-Transforming Encoder-Decoder (ST-ED) framework, that encodes both task-relevant factors, such as gaze and head orientation, and task-irrelevant factors. This model is designed to process image pairs, enabling it to effectively distinguish subject-specific embeddings from those associated with gaze transformations. This crucial differentiation allows the model to leverage pseudo labels for learning transformation-related attributes with minimal calibration samples.

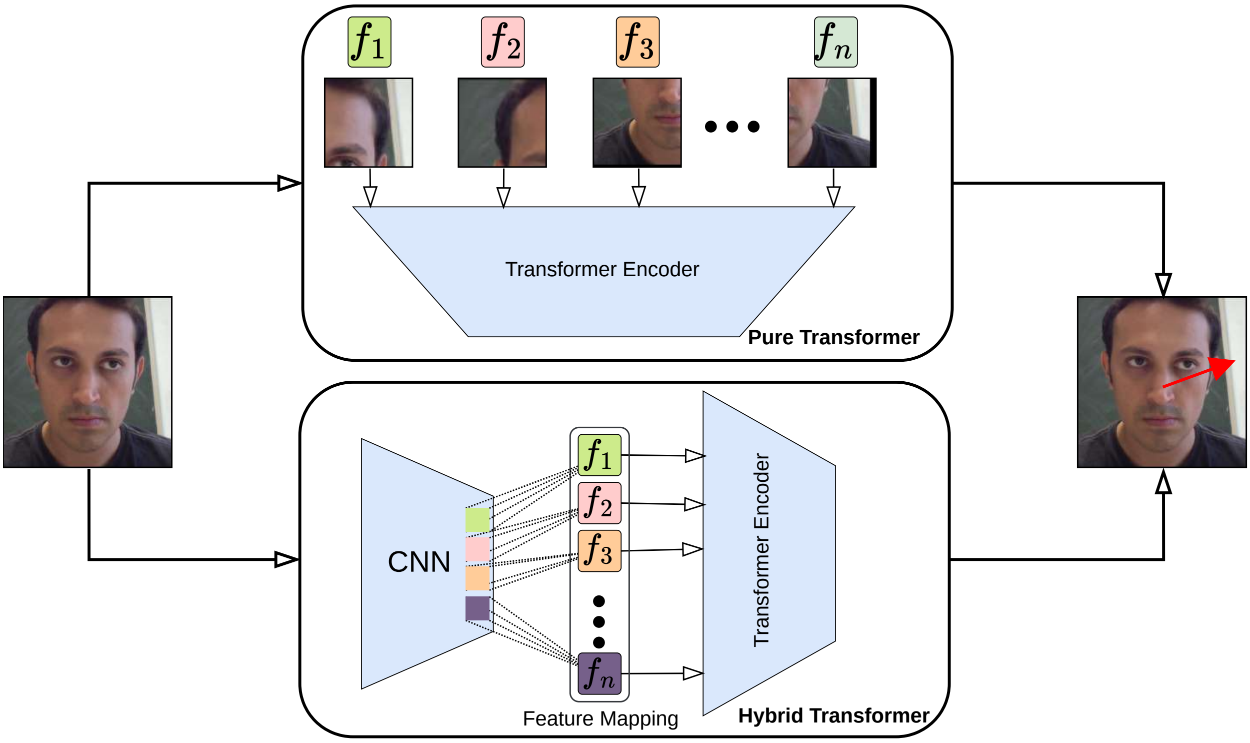


Figure 2.10: The transformer model developed by Cheng et al [31]. GazeTR-Pure, with a pure transformer architecture designed for extracting global gaze features, alongside GazeTR-Hybrid a hybrid version combining CNNs and transformers for gaze direction prediction.

Gaze-Redirection-Net [232], leverages a pre-trained gaze redirection network for unsupervised learning of gaze representations. The model aims to generalize eye appearance across different gaze directions without explicit supervision. This model focuses on learning a generic representation of the eye highlights the potential of unsupervised learning in enhancing gaze estimation models, particularly in terms of representation learning and gaze manipulation.
RITnet [25], tackles the segmentation of the eye region, an essential preliminary step for accurate gaze estimation. The model introduce a hybrid model that combines elements of U-Net and DenseNet within a Fully Convolutional Network (FCN) to efficiently segment the eye. This approach provides a solid foundation for subsequent gaze estimation efforts and is designed to balance the trade-off between performance and computational complexity.

4) Transformer-based gaze estimation models: The advent of transformer models in computer vision has led to their exploration in the field of gaze estimation, offering new methodologies beyond traditional CNNs. Specifically, the vision transformer (ViT) framework has been adapted to gaze estimation tasks, showing two distinct approaches: GazeTR-Pure and GazeTR-Hybrid as shown in Figure 2.10.
GazeTR-Pure [32], uses a pure transformer architecture for gaze estimation, which

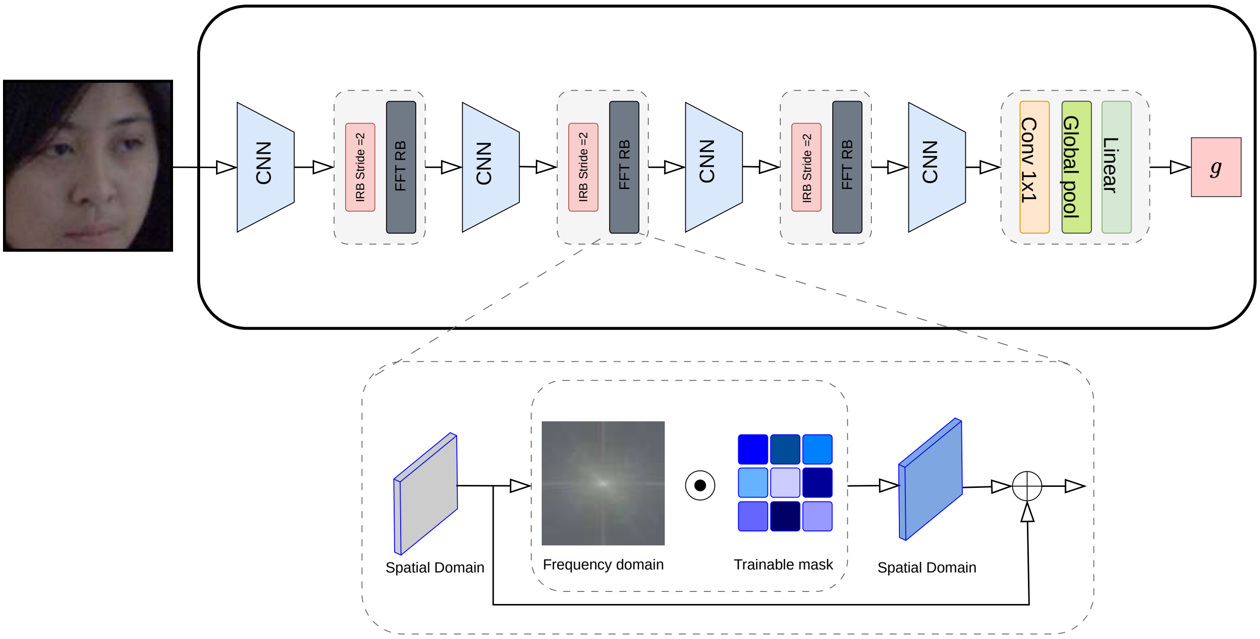


Figure 2.11: The lightweight model developed by Xu et al [222], namely FR-Net, that enhances gaze estimation by integrating Fast Fourier Transform (FFT). The FFT Residual Blocks are utilized to extract both depth spatial domain and frequency domain features, which are then fused to form the input for the subsequent layer.

directly processes facial images. The framework introduce a novel mechanism that learns to capture global image feature, enabling the model to analyze facial features from a comprehensive perspective instead of focusing on local patches. This approach highlights the significant potential of transformers to improve the accuracy of gaze estimation through a holistic understanding of the images.

GazeTR-Hybrid [32], addresses the challenges associated with using pure transformers for regression tasks like gaze estimation. This model introduces a hybrid approach that combines CNNs for extracting local features with transformers for synthesizing global information. This hybrid approach leverages the strengths of both architectures: CNNs for capturing detailed local features and transformers for understanding the overall context. The GazeTR hybrid model not only enhances the accuracy of gaze estimation but also reduce the computational costs associated with the process.

SAtten-Net [135], integrate self-attention with convolution and deconvolution to enhance eye gaze estimation from full-face images. They efficiently model local contexts and reduce the computational costs of applying self-attention directly to images, while also ensuring that crucial spatial information is retained through a deconvolution process.

5) Lightweight-based gaze estimation models: Lightweight networks play a key role in the development of efficient network architectures for various domains. Starting with Xception network [37, 240] that adapts deep separable convolutional layers to significantly reduce the number of parameters and enhancing the efficiency of network designs. Furthermore, MobileNet [74] advanced this approach by employing deep separable convolutional layers, moving away from traditional models reduction methods. Further improvements were made with MobileNetV2 [161], which introduced backward residuals and linear bottlenecks, optimizing the MobileNet architecture for better accuracy and lower computational demand. Furthermore, the MobileViT [61, 125, 126, 191] series combines the strengths of CNNs and Transformers, providing exceptional performance on a wide range of mobile vision tasks. These lightweight networks opened new avenues in the field of effective gaze estimation.

GazeCaps [194], utilizes self-attention-routed capsules to represent various facial properties crucial for gaze estimation. the model encapsulates the non-linear transformations of facial features with head movement into distinct capsules. These capsules are sensitive to transformations, providing a vectorial representation that effectively captures the complex changes in facial components related to gaze direction. Furthermore, the framework incorporates a Self-Attention Routing (SAR) module that dynamically focuses on relevant capsules containing significant information for gaze estimation. This mechanism allows for optimization in a single process without the need for iterations.

EM-Gaze [261], incorporates eye context correlation and metric learning to enhance the accuracy of 2D gaze estimation on mobile devices. The framework is designed around two-stream collaborative framework, integrating an attention-based module to correlate and fuse contextual features from the left and right eyes. This approach leverages metric learning for gaze classification across display quadrants, thereby enhancing both regression and classification performance. The framework sets a new benchmark in gaze estimation, illustrating its potential for practical applications in human-computer interaction.

I2DNet [133], leverages dilated convolutions and a unique differential layer to improve real-time, subject-independent gaze accuracy. Dilated convolutions enhance the network's receptive field without compromising spatial details, while the differential layer focuses on gaze-relevant features by emphasizing differences between left and right eye features. Through a user study involving a 9-block pointing and selection task, I2DNet demonstrated improved performance in selection time and reduced miss selections, highlighting its potential for creating efficient gaze-controlled interfaces using standard webcams.

FR-Net [222], uses Fast Fourier Transform (FFT) to extract gaze-relevant features within frequency domains, thereby diminishing both the parameter count and computational demands. A notable innovation is the introduction of a shortcut component that emphasizes spatial domain analysis as shown in Figure 2.11. Experimental results underscore the superior performance of FR-Net in achieving lower gaze error angles, while employing considerably fewer parameters and FLOPs in comparison to existing state-of-the-art gaze estimation methods.

2.4.2 Validation

This section reviews the commonly followed evaluation procedures used in gaze estimation, including various datasets and the metrics adopted for assessing performance. It emphasizes the diversity of datasets and the benchmarks set for comparison, providing a comprehensive overview of how performance in gaze estimation is measured and benchmarked across different studies.

2.4.2.1 Datasets

The advancement of gaze estimation research is significantly influenced by the development and availability of diverse datasets, which have evolved from constrained laboratory settings to more complex and natural indoor and outdoor environments. As illustrated in Table 2.1, gaze estimation datasets vary across several characteristics, including the number of subjects, head pose, gaze, total images or videos, and the application environment. This evolution in dataset collection techniques reflects the growing need for models that can operate under varied and less controlled conditions, thereby pushing the boundaries of gaze estimation technology. A summary of the important datasets is presented as follows:

EyeDiap [55]: is designed to address the lack of standardized benchmarks for evaluating gaze estimation algorithms. It was created to assess the robustness of gaze estimation methods across various conditions, including head pose variations, individual differences among participants, changes in lighting conditions, and different target types (screen or 3D objects). This dataset provides a rich source of both RGB and RGB-D data that support comprehensive evaluations with its systematic recording methodologies and pre-processed data. This dataset facilitates the development of more accurate and adaptable gaze estimation methods.
TabletGaze [86]: provides an essential resource for addressing the challenges of gaze estimation on tablets. It comprises data from 51 subjects engaging in a variety of natural user interactions across 35 gaze locations. This dataset is notable for its inclusion of real-world variables, such as different user postures, the presence of eyeglasses, and diverse lighting conditions that significantly influence gaze tracking accuracy. It encompasses a range of postures including standing, sitting, slouching, and lying. The inclusion of various postures adds complexity and realism to the dataset making it a valuable tool for developing more adaptive gaze estimation systems.
UTMultiview [171]: is designed for the purpose of learning a person and pose independent gaze regression function. It contains a large collection of gaze data, with 64,000 images from 50 subjects across 160 gaze directions and 8 head poses. This dataset stands out due to its comprehensive calibration and annotation in a 3D world coordinate system, ensuring consistent and accurate position data across all subjects. The extensive data collected, including a wide range of head poses and gaze directions,

Table 2.1: A comparison of gaze estimation datasets across various characteristics: including the number of subjects, head pose, gaze, total images or videos and application environment.

Dataset	Sub	Head Pose	Gaze	Total	Env
ColumbiaGaze [168]	58	0°,±30 °	±15°±10°	6K Images	Indoor
UTMultiview [171]	50	±36°, ±36°	±50°, ±36°	1.1M Images	Indoor
EyeDiap [55]	16	±15°, ±30°	±25°, ±20°	94 Videos	Indoor
MPIIGaze [252]	15	±15°, ±30°	±20°, ±20°	45k Images	Indoor
GazeCapture [107]	1,474	±30°, ±40°	±20°, ±20°	2.4M Images	Hybrid
MPIIFaceGaze [250]	15	±15°, ±30°	±20°, ±20°	45K Images	Indoor
InvisibleEye [178]	17	Unknown	2560X1600	280K Images	Indoor
TabletGaze [86]	51	±50°, ±50°	±20°, ±20°	816 Videos	Indoor
RT-Gene [54]	15	±40°, ±40°	±40°, -40°	123K Images	Indoor
Gaze360 [96]	238	±90°, u/k	±140°, -50°	172K Images	Hybrid
NVGaze [101]	30	Unknown	30°X 40°VF	4.5M Images	Hybrid
RT-BENE [40]	17	±40°, ±40°	±40°, -40°	243K Images	Indoor
ETH-XGaze [243]	110	±80°, ±80°	±120°, ±70°	1.1M Images	Indoor
EVE [143]	54	±80°, ±80°	±80°, ±80°	4.2K Videos	Indoor
GazeFlow [156]	130,339	Variable	Variable	122K Images	Hybrid
UnityEyes [211]	N.A	Variable	Variable	1M Images	Sync
HUST LEBW [76]	172	Variable	Variable	673 Videos	Hybrid
VACATION [52]	206,774	Variable	Variable	97K Images	Hybrid
mEBAL [42]	38	Variable	Variable	756K Images	Indoor
LAEO [105]	485	Variable	Variable	800K Images	Hybrid
GOO [177]	100	Variable	Variable	201K Images	Hybrid
OpenNEEDS [51]	44	Variable	Variable	2M Images	Hybrid

supports the synthesis of new appearances and the training of a gaze estimation model that outperforms existing methods. The dataset's unique design allows for detailed 3D reconstruction of eye regions, which enhances the generation of synthesized training data, crucial for improving the accuracy of gaze estimation algorithms.

MPIIFaceGaze [248]: encompasses 213,659 images from 15 participants that capture a wide array of real-life scenarios. It includes varying illuminations and environments experienced during everyday laptop use over three months. This dataset offers a level of diversity in appearance, head poses, and gaze directions. The data are collected under controlled laboratory conditions, providing a rich representation of real-world conditions. The MPIIGaze dataset fills a critical gap in gaze estimation research by offering a comprehensive resource that significantly enhances the ability to develop and test gaze estimation models in scenarios that closely mimic everyday settings.

GazeCapture [107]: marks a significant step forward in the domain of eye tracking, enabling eye tracking technologies on commonly used devices such as smartphones and tablets without the need for additional sensors. As the first dataset of its kind, it comprises nearly 2.5 million frames collected from over 1,450 participants. This vast collection captures a wide array of participant backgrounds, lighting conditions, and head movements. It greatly enrich the data available for enhancing the accuracy and generalizability of eye tracking models. The scale and variety offered by the GazeCapture dataset represent a pivotal resource for advancing eye tracking technologies.

Gaze360 [95]: stands out as the largest of its kind for 3D gaze estimation, featuring 172K images from 238 subjects of different genders, ages, and ethnicities. It encompasses a wide range of head poses and distances across various indoor and outdoor settings. This diversity and scale enable the dataset to significantly advance the study and application of gaze estimation in unconstrained environments. The images were captured using a Ladybug multi-camera system in various settings, presenting a substantial test for unconstrained gaze estimation models. By providing such a comprehensive dataset, researchers address the critical need for large and diverse annotated training data, which has been a limiting factor in the field. The Gaze360 dataset supports cross-dataset evaluations and domain adaptation research, marking a notable contribution to gaze estimation research and its practical applications.

RT-GENE [54]: comprises 122,000 labeled images and 154,755 unlabeled images from 15 participants. It addresses the challenge of accurately estimating gaze over large distances and across a broad range of head poses and eye gaze angles. It features a comprehensive collection of images with ground-truth annotations for both gaze and head poses. These annotations are obtained using a motion capture system for head poses and mobile eye-tracking glasses for eye gaze. A novel aspect of this dataset is the use of a semantic in-painting technique to replace the appearance of the eye-tracking glasses with realistic skin texture in the images. This ensures the dataset's relevance for both training and testing, enhancing its utility for developing more accurate gaze estimation models.

ETH-XGaze [243]: provides high-resolution images from over one million images collected from 110 participants. It designed to enhance gaze estimation technologies under extreme variations in head pose and gaze direction. It captured using a custom setup with 18 digital SLR cameras and adjustable lighting including a diverse group of subjects wearing varied eye wear. It encompasses a wide array of head poses and gaze directions annotated with high-quality ground truth targets. These features aim to significantly improve the robustness of gaze estimation methods across different conditions and establish a standardized evaluation protocol to unify future research in this field.

The transition from constrained to more unconstrained, real-world settings in gaze estimation datasets reflects the community's response to the need for models that are robust across various operational conditions. With each new dataset, the gaze estimation field moves closer to developing systems that can perform reliably in the diverse settings encountered in daily life, from personal device interaction to automotive and accessibility applications.

2.4.3 Evaluation Metric

There are two types of evaluations commonly used in gaze estimation: Within-dataset Evaluation and Cross-dataset Evaluation. Within-dataset Evaluation assesses the model's performance on unseen subjects from the same dataset. The dataset is divided into a training set and a test set based on the subjects, ensuring no overlap of subjects between the training and test sets. Most gaze datasets provide a predefined within-dataset evaluation protocol that pre-divides the data into training and testing sets. Cross-dataset Evaluation assesses the model's performance in unseen environments. The model is trained on one dataset and tested on another, evaluating its ability to generalize to new conditions and datasets. Accuracy or error in gaze estimation is quantified in terms of angular error (in degrees) and gaze location (in pixels or cm/mm). Angular error calculates the deviation between the actual and predicted gaze directions, while Euclidean distance measures the difference between the original and predicted points of gaze (PoG).

Angular error: this metric is typically used to assess the accuracy of 3D gaze estimation methods. It involves computing the angle between the actual gaze direction $\mathbf{g}$ and the estimated gaze direction $\hat{\mathbf{g}}$, both represented in 3D space. The angular error is defined as the measure of accuracy in terms of the angular difference between the actual and estimated gaze directions. Assuming that the ground truth gaze direction is $\mathbf{g} \in \mathbb{R}^3$ and the predicted gaze vector is $\hat{\mathbf{g}} \in \mathbb{R}^3$, the gaze angular error (in degrees) is calculated by:

$$\mathcal{L}_{\text{angular}} = \arccos\left(\frac{\mathbf{g} \cdot \hat{\mathbf{g}}}{\|\mathbf{g}\| \|\hat{\mathbf{g}}\|}\right) \tag{2.1}$$

Euclidean Distance: used to evaluate the accuracy of 2D gaze estimation methods. It measures the straight-line distance between the actual gaze position $\mathbf{P}$ and the estimated gaze position $\hat{\mathbf{P}}$, both in a 2D space. This metric serves as an accuracy measure by quantifying the spatial discrepancy between the actual and predicted gaze points. Assuming that the P_x, P_y are the coordinates of the actual gaze position, and

$\hat{P}_x, \hat{P}_y$ are the coordinates of the estimated gaze position, the gaze euclidean distance is calculated by:

$$\mathcal{L}_{\text{Euclidean}} = \sqrt{(\hat{P}_x - P_x)^2 + (\hat{P}_y - P_y)^2} \tag{2.2}$$

2.4.4 Adaptation

In the realm of gaze estimation technology, achieving accurate gaze direction prediction from images has been a focal point, with various approaches demonstrating promising results within dataset evaluations. However, the application of these models to new, unseen domains (cross-dataset evaluations) often faces challenges due to factors such as variations in face appearance, head pose, image quality, and illumination. To tackle the issue of gaze generalization to new environments, two primary strategies are employed: domain generalization and domain adaptation.

2.4.4.1 Domain generalization

Domain generalization techniques strive to enhance model performance on unseen domains by leveraging only the data from source domains, without incorporating any labels from the target domain.

PureGaze [29], introduces a domain generalization framework for gaze estimation designed to purify gaze features by removing gaze-irrelevant factors such as illumination and identity, without the need for target domain data. This approach aims to enhance cross-dataset performance using a self-adversarial framework to purify the extracted features. The framework encompasses two key tasks: gaze estimation to preserve gaze-relevant information and adversarial reconstruction to eliminate general image information. The implementation utilizes two shared weight backbones for feature extraction, a two-layer MLP for gaze estimation, and an SA-Module for image reconstruction, which facilitates feature purification through adversarial learning. PureGaze significantly enhances existing gaze estimation methods without requiring additional inference parameters or training images. This proves its effectiveness and plug-and-play capability for improving generalization in gaze estimation across different domains.
SAtten-Net [135], proposes an advanced framework for eye gaze estimation that combines self-attention with convolution and deconvolution, aiming to enhance generalization performance beyond traditional methods. This approach leverages self-attention to capture the global context, significantly boosting the model's generalization capabilities. By incorporating convolution projection and deconvolution both before and after the self-attention process, the model efficiently balances between modeling local contexts and reducing the computational demands associated with self-attention. This strategy ensures an effective equilibrium between accurately modeling gaze behavior

and maintaining computational efficiency. These methods show potential in extending to new domains; however, their effectiveness is limited by the scarcity of annotated in-the-wild gaze datasets that can encompass the wide array of environmental conditions present in real-world settings.

2.4.4.2 Domain adaptation

Domain adaptation approaches utilize data from the target domain during the training process to directly address domain gap.

GazeAdv [197], addresses the challenge of generalizing eye tracking across different subjects and environments by developing a novel Bayesian adversarial learning framework. This approach systematically tackles the factors that hinder generalization in appearance-based gaze estimation, such as variations in appearance, head pose, and overfitting issues commonly associated with point estimation. By integrating adversarial learning with Bayesian inference, GazeAdv learns features that are responsive to gaze direction while being robust to variations in appearance and pose. Additionally, by adopting a Bayesian framework, the model performs gaze estimation using a distribution of parameters rather than a single set of parameters, which effectively mitigates overfitting.
RUDA framework [12], proposes the Rotation-enhanced Unsupervised Domain Adaptation (RUDA) framework, which leverages the rotation consistency property of gaze estimation. It assigns sub-labels to target domain images based on relative rotation angles by rotating original images at various angles for training, and applies domain adaptation under rotation consistency constraints. This innovative approach enables effective domain adaptation without requiring real labels in the target domain. The RUDA framework is evaluated across four cross-domain gaze estimation tasks, demonstrating significant performance improvements.
PnP-GA framework [116], The framework is distinguished by its ensemble approach, where networks collaborate under the guidance of outliers to adapt pre-existing gaze estimation algorithms to novel environments or individuals. This outlier-guided collaborative learning strategy significantly improves upon traditional unsupervised domain adaptation methods, showing substantial performance enhancements in various domain adaptation tasks. The ability of the PnP-GA framework to seamlessly integrate with existing gaze estimation networks without architectural modifications highlights its practicality for real-world applications. This capability sets a new benchmark in the field of gaze domain adaptation.
CRGA framework [203], extract stable representations from source domain using contrastive domain generalization (CDG) module. Additionally, it learn from pseudo labels in the target domain using contrastive self-training adaptation (CSA) module. The core innovation of CRGA lies in the Contrastive Regression loss, which is designed

to bring features with similar gaze directions closer together and push dissimilar ones apart, thereby enhancing the model's adaptability to new domains without requiring labeled data. The effectiveness of CRGA has been demonstrated across several domain adaptation tasks, significantly improving performance over baseline models and surpassing current state-of-the-art domain adaptation methods in gaze estimation. This breakthrough paves the way for more robust gaze estimation models capable of adapting to varied environments and subjects without the need for extensive retraining or labeled target domain data.

While these methods offer improvements in cross-dataset performance, they typically rely on extra models or annotations, which might not always be accessible in practical scenarios. Moreover, the resulting models are fine-tuned for specific domains rather than being universally applicable to any unseen domain, highlighting a limitation in their generalizability.

3 Improving Gaze Estimation Accuracy Using Combined Classification And Regression Losses

3.1 Introduction

Gaze estimation technology has gained significant advances in the last decade. This growing interest is driven by its wide array of applications, which include human-robot interaction [2, 71, 169], driver engagement [1, 59, 79, 139, 190], and augmented reality [39, 205]. The ability to accurately predict a person's gaze direction can significantly enhance the interface between humans and machines, offering more intuitive and efficient interactions. Gaze estimation methods can be broadly categorized into conventional and deep learning-based approaches. Conventional-based methods utilize geometric models and specific eye features to determine gaze direction. However, these methods face challenges such as variability in eye anatomy, lighting conditions, occlusions, head movements, and environmental changes.

The advent of deep learning and CNNs has revolutionized the field of computer vision [130, 138, 226, 227, 238], demonstrating remarkable performance across a diverse array of tasks such as image recognition [47, 98], object detection [257, 271], and semantic segmentation [134, 134]. Early methods [27, 36] focus on using both eye and face images for gaze estimation, since eyes directly indicate gaze direction and the facial image offers critical cues such as head pose information. However, the need for multiple backbones to analyze both eye and face images significantly increased computational demands. Recent approaches have shown promising results by using the entire facial image that contains information about the eye movement and head pose, which is important for gaze estimation. One of the most important challenges for these approaches is the need to capture fine-grained features of the eye appearance while simultaneously filtering out irrelevant facial features that do not contribute to gaze estimation. Further, the eye region, a relatively small area of the overall facial image, contains essential cues that are necessary to accurately predict gaze direction. Researchers have developed various CNN architectures [10, 27, 36, 54, 107, 107, 113, 146, 204, 233, 256] to overcome these limitations, utilizing different backbones and specially designed networks to improve the accuracy of gaze estimation. However, CNNs often struggle to extract fine-grained features, since they are prone to focusing on larger, more continuous areas of the face, potentially neglecting the subtle details critical for accurate gaze estimation.

This chapter proposes a different strategy to address the challenge of improving fine-grained gaze features for accurate gaze estimation. Traditional methods have predominantly adapted regression loss functions like mean square error (MSE)[14, 27, 54] and mean absolute error (MAE)[31, 242, 253] to penalize network learning. However, this can lead to difficulties in emphasizing critical eye regions in the face image which are vital for accurate gaze information, thereby negatively affecting the overall performance. Additionally, these methods predict gaze angles (pitch and yaw) directly using a separate fully connected regression layer. The utilization of this methodology may impose limitations on the model's ability to deal with the intricate relationships and varied patterns present in gaze data. Additionally, simultaneous prediction of both angles may limit the network's ability to recognize the various features and variations unique to each gaze angle, hence affecting its performance.

In this chapter, L2CS-Net is proposed as an effective method to accurately capture fine-grained gaze features from tiny eye regions in facial images. L2CS-Net leverages the strength of a dual-branch CNN and combines classification and regression losses to enhance both the accuracy and robustness of gaze estimations. L2CS-Net introduces a dual-branch CNN architecture that separates fully connected layers to predict the pitch and yaw of each gaze angle. This design enables explicit learning of independent features and emphasizes the distinct information associated with each angle, thereby enhancing feature separability. Furthermore, L2CS-Net employs combined classification and regression losses for each separate angle prediction. This joint optimization of classification and regression objectives facilitates the extraction of more informative gaze features. L2CS-Net utilizes a softmax layer along with a cross-entropy loss to obtain the coarse gaze direction (gaze direction within specific angular intervals). In contrast, fine-grained predictions are achieved by calculating the expectation of the gaze-bin probabilities followed by a gaze regression loss. Through extensive experiments on various challenging datasets, L2CS-Net has improved overall performance and outperforms existing state-of-the-art methods in terms of accuracy.

3.2 Key Contributions

As shown in Fig. 5.1, this chapter address the critical challenge of accuracy in gaze estimation by capturing fine-grained features from the tiny eye region within facial images. The key contributions of this chapter are presented in detail:

- L2CS-Net: A novel two-branch CNN architecture that utilizes two separate fully connected layers dedicated to each gaze angel prediction. This separation allows for the explicit and focused learning of discriminative features pertinent to each gaze angle, thereby enhancing the model's ability to accurately capture the detailed variations associated with different gaze directions.

- Innovative Multi-Loss Function: Building upon the architectural, L2CS-Net introduces a multi-loss function that combines classification and regression losses.

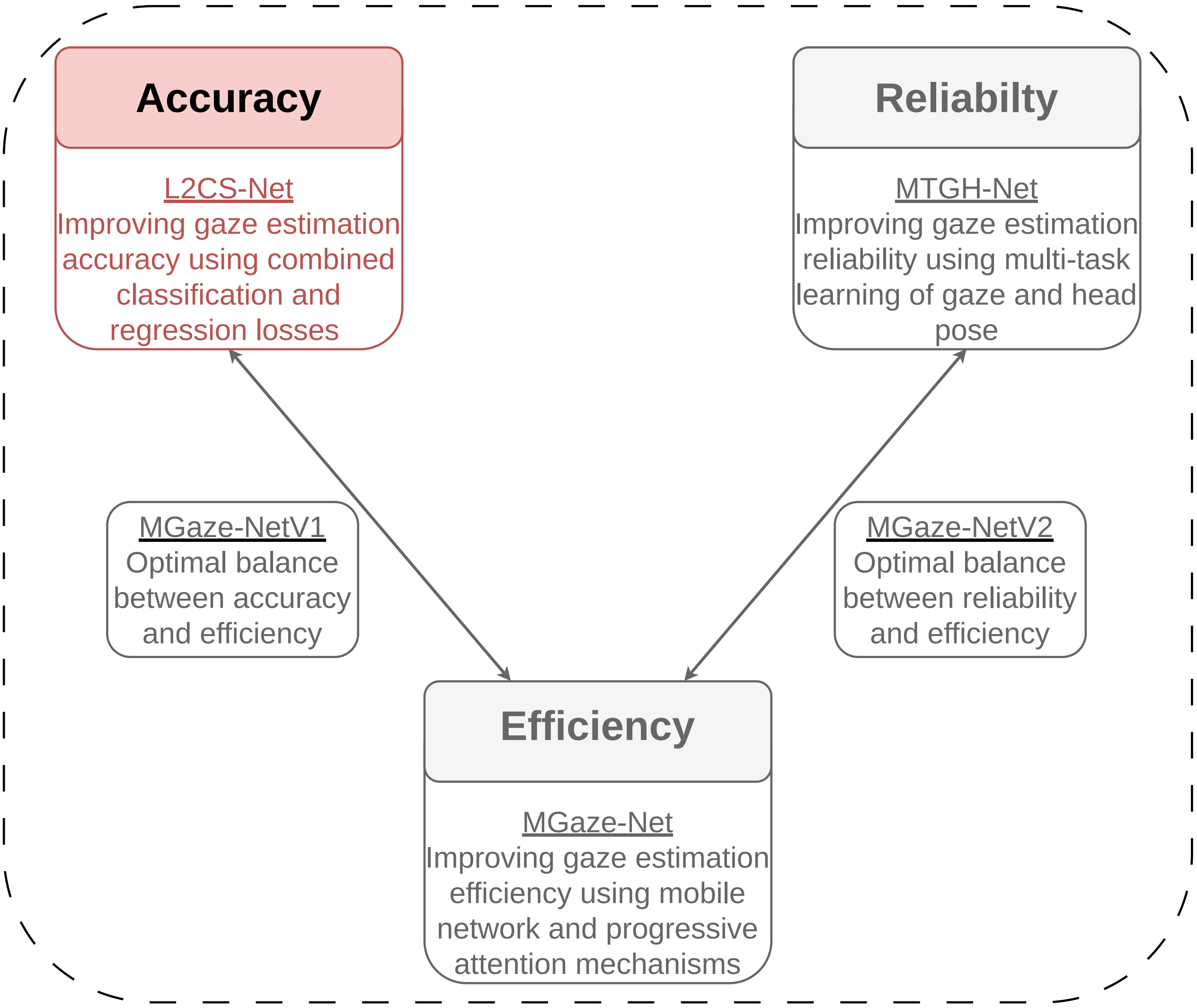

Accuracy
L2CS-Net
Improving gaze estimation accuracy using combined classification and regression losses
Reliabilty
MTGH-Net
Improving gaze estimation reliability using multi-task learning of gaze and head pose
MGaze-NetV1
Optimal balance between accuracy and efficiency
MGaze-NetV2
Optimal balance between reliability and efficiency
Efficiency
MGaze-Net
Improving gaze estimation efficiency using mobile network and progressive attention mechanisms

Figure 3.1: Improving gaze estimation accuracy using combined classification and regression losses. The contribution of this chapter is the design of an accurate gaze estimation model that can extract fine-grained gaze features from facial images.

This hybrid approach enables L2CS-Net to leverage the strengths of both classification, for coarse gaze direction prediction, and regression, for refining these predictions to improve fine-grained features. The dual nature of this loss function facilitates a comprehensive optimization process, ensuring that the network not only classifies gaze direction within predefined angular bins but also fine-tunes these estimations to closely match the actual gaze direction.

- Computational Efficiency: L2CS-Net utilizes EfficientNetV2, an efficient CNN backbone with its optimal balance between accuracy and computational efficiency. This design choice help to achieve accurate gaze estimation while minimizing computational demands.

- Extensive Validation on Challenging Datasets: The accuracy and robustness of the proposed L2CS-Net are rigorously validated through extensive experiments on four challenging datasets: MPIIFaceGaze, GazeCapture, Gaze360, and RT-Gene. The comprehensive evaluation demonstrates that L2CS-Net outperforms existing state-of-the-art methods by up to 18% across these datasets, achieving new benchmarks in gaze estimation accuracy.

- Ablation Study: Through a detailed ablation study, the contributions of each component of the proposed L2CS-Net are analyzed. These studies reveal the critical role of the two-branch architecture and the multi-loss function in enhancing gaze estimation accuracy. The effectiveness of these components is empirically validated by systematically varying the network architecture and loss functions.

The remainder of this chapter is organized as follows. Section 3.3 presents the relevant prior works. Section 3.4 presents the details of the proposed L2CS-Net. In Section 3.5, experimental, quantitative and quantitative analysis of L2CS-Net is presented. Finally, Section 3.6 contains the conclusion.

3.3 Prior Work

Accurate gaze estimation has prompted extensive research, leading to the development of diverse methodologies classified primarily into model-based, feature-based and appearance-based techniques [8, 89]. Model-based methods utilize the anatomy of the eyeball, primarily through two types of models: elliptical iris boundary models [20, 179, 213] and spherical eyeball models [8, 89]. However, issues such as partial eyelid closure pose significant challenges in accurately extracting the limbal ellipse from images, and uncertainties in determining the 3D center of the eyeball continue to affect the accuracy of gaze estimation. Feature-based gaze estimation techniques [83, 254, 255] regress gaze direction after identifying specific key points within the input image of the eye. However, they face challenges in accurately localizing features under diverse conditions, including occlusions and extreme head movements. In contrast, appearance-based methods offer a more versatile approach by focusing on a holistic analysis of the entire eye image, avoiding issues with specific feature tracking. The accuracy of these techniques is often compromised by variations in individual appearances, head poses, lighting conditions, and environmental settings.

Convolutional Neural Networks (CNNs) [78, 110, 202] have emerged as a powerful tool in this domain. CNNs demonstrate remarkable proficiency in learning complex mappings from facial or eye images to gaze direction. The early approaches use conventional CNN backbones such as LeNet [248], AlexNet [251], and VGG [54] to extract gaze features from images. Zhang et al. [251] introduce a VGG CNN-based architecture, marking a pivotal shift to using CNNs for gaze estimation from single-eye images [253]. The enhanced model incorporated spatial weights to prioritize facial regions critical for gaze information, optimizing the extraction of relevant features [251]. Krafka et al.[107] propose a network utilizing multiple data channels including eye images, face images, and face grid information to robustly capture essential gaze cues [107]. Fischer et al. [54] propose a gaze estimation method that incorporates the head pose vector and features extracted using VGG-Net from eye crops to predict accurate gaze angles.

Subsequent innovations introduced specially designed CNN architectures tailored for gaze estimation tasks, aiming to enhance the extraction of fine-grained features from the eye region. CA-Net [30] adapts a coarse-to-fine strategy equipped with an attention module to capture precise fine-grained gaze features. Cheng et al. [36] proposed FAR-Net, which leverages the two-eye asymmetry property to estimate 3D gaze angles. They assign asymmetric weights to the loss functions of each eye and aggregate them for improved performance. Wang et al. [201] adopted ID-ResNet that contains a residual neural network structure with an embedding layer of personal identity to overcome the different geometric parameters between subjects. Despite these advancements, a common challenge across conventional CNN-based methods is their tendency to focus on larger, more contiguous areas, overshadowing the subtle yet crucial gaze related features.

Recognizing these challenges, recent studies have explored various strategies to refine the CNN's ability to capture and analyze fine-grained features crucial for gaze estimation. Techniques such as temporal models [95, 140, 142, 196, 262], attention mechanisms.[32, 32, 135], generative models [144, 232, 258], and multitask learning [105, 113] have been proposed to direct the focus of CNNs towards relevant features of the eye region while filtering out irrelevant information. However, the search for an optimal balance between accuracy, computational efficiency, and adaptability to unconstrained settings remains ongoing.

3.4 Framework

This section presents the comprehensive framework of L2CS-Net, a two-branch multi-loss CNN architecture designed for fine-grained gaze estimation. L2CS-Net is an integration of a two-branch CNN architecture and a loss function design of classification and regression losses. This section outlines the architecture, the proposed loss function, and the rationale behind each component, illustrating how they collectively contribute to improving gaze estimation accuracy from face images. A complete pipeline of the proposed framework is depicted in Figure 3.2.

3.4.1 Network architecture

The core of L2CS-Net is the two-branch CNN architecture, specifically tailored to address the inherent challenges of gaze estimation from facial images. Initially, the network employs EfficientNetV2 [175] as a shared feature extraction backbone to process the input facial images. This backbone is chosen for its ability to efficiently

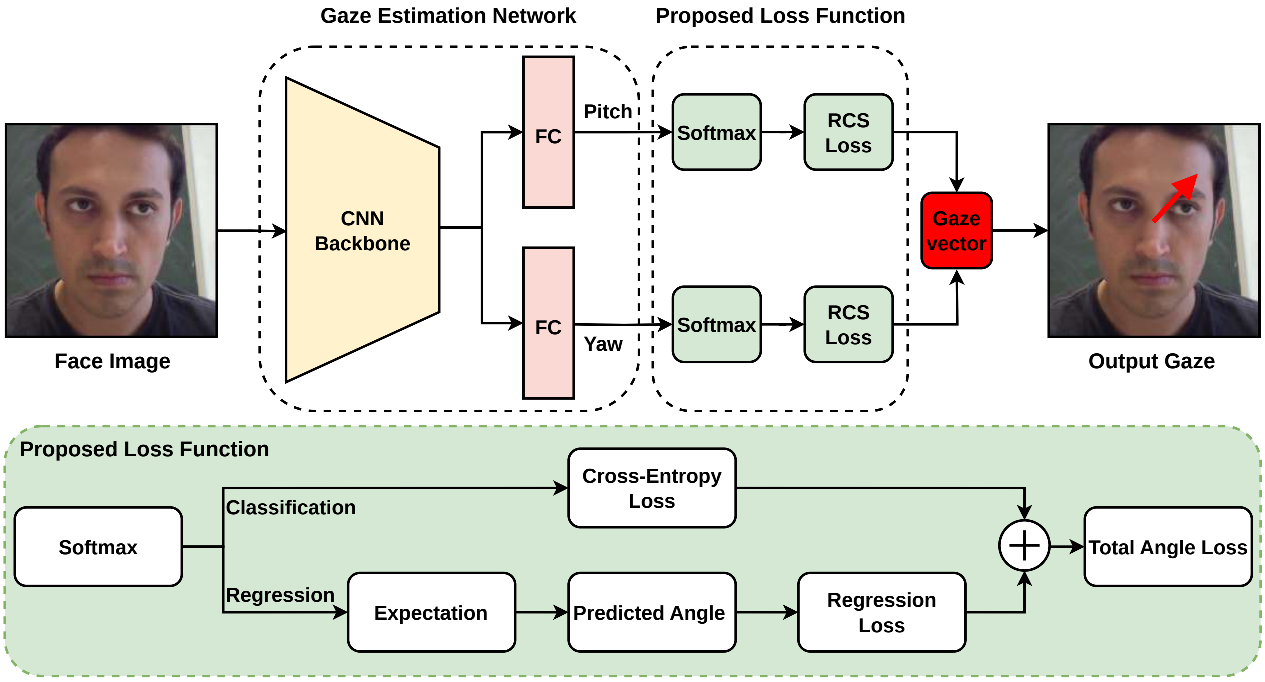


Figure 3.2: L2CS-Net with two branch CNN architecture and two combined regression classification losses. A reliable softmax with cross-entropy loss for classification and the expectation of the gaze bin probabilities with mean absolute error for regression.

capture a wide range of features from the input images, providing a rich feature set for the subsequent branches. Further, it utilizes compound scaling and a streamlined structure to minimize computational demands while maintaining high performance. The initial shared backbone ensures that a comprehensive feature set is available, while the subsequent branches refine these features to produce accurate gaze angle predictions.

L2CS-Net is designed to output separate predictions for pitch and yaw gaze angles, leveraging the inherent distinctions between these two components of gaze direction. The dual-branch design allows the network to optimize the feature extraction process for each angle, leading to more accurate predictions. Each branch consists of a fully connected layer dedicated to either pitch or yaw, allowing the network to learn and emphasize discriminative features pertinent to each angle independently. The design choice to separate prediction tasks into two branches stems from the observation that pitch and yaw angles encapsulate different characteristics of gaze direction and may exhibit varied patterns in facial images. By dedicating each branch in pitch or yaw estimation, the network can develop a more nuanced understanding of gaze-related features, leading to improved accuracy and robustness in gaze prediction.

To align the network with the proposed multi-loss function, the output of each fully connected layer is adapted to predict various output classes or bins determined by the angular range specific to each dataset and a uniform bin width. This ensures comprehensive coverage across the range of gaze angles. Through the application of a softmax layer, the output logits are transformed into a probability distribution for each bin, with the innovative multi-loss function further refining network performance. This integration of softmax and cross-entropy losses, complemented by regression loss, provides a robust framework for accurate bin classification and continuous-angle prediction, promoting the model's stability and accuracy.

3.4.2 RCS Function

A key innovation of L2CS-Net is the introduction of a multi-loss function that synergizes classification and regression losses, named the Regression Classification Loss (RCS). This composite loss function is pivotal in refining the gaze estimation process, ensuring both coarse and fine-grained accuracy in the predictions.

3.4.2.1 Classification Loss

The RCS classification component focuses on categorizing the gaze direction into discrete bins, each representing a specific range of angles. This categorization aids in capturing the coarse gaze direction, providing a preliminary estimate that narrows down the possible range of the true gaze angle. The classification loss encourages the network to accurately assign probabilities to these bins, facilitating a rough and crucial localization of gaze direction. A suitable classification framework for the gaze estimation problem is multi-class classification. For the gaze classification part, softmax layer is employed followed by categorical cross-entropy loss. The softmax layer used to transform the network's bin logits into probabilities for each gaze bin as detailed below:

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{b} e^{x_j}} \tag{3.1}$$

where $f(x_i)$ is the softmax output value, x_i is the logits of bin i and b is the number of output bins. Following, the cross-entropy loss used to estimate the error between the output bin probabilities and gaze targets as follows:

$$CE = -\sum_{i=1}^{b} t_i \log f(x_i) \tag{3.2}$$

where CE is the cross-entropy loss, t_i and $f(x_i)$ are the target ground truth and the softmax probability value for each bin i in b. L2CS-Net adapt one-hot encoding gaze targets, so only the positive bin i_p keeps its value in the loss calculation and discards the elements of the summation which have zero values. The final gaze classification loss $\mathcal{L}_{cls}$ is estimated as follows:

$$\mathcal{L}_{cls} = -\log\left(\frac{e^{x_{i,p}}}{\sum_{j=1}^{b} e^{x_j}}\right) \tag{3.3}$$

where $x_{i,p}$ is the output probability for the positive bin.

3.4.2.2 Regression Loss

In contrast to the classification component, the regression loss fine-tunes the gaze direction estimate within the specified bin range. Using regression, the network refines its predictions to achieve a level of accuracy that closely matches the actual gaze direction. This component of the loss function ensures that the model not only identifies the correct bin for gaze direction but also accurately estimates the gaze angle within that bin, capturing the fine-grained details necessary for accurate gaze estimation.

To output continuous angle values from discrete bin outputs, L2CS-Net involves a two-stage process. Initially, a softmax layer is applied to the fully connected layer outputs, converting the network logits into probabilities for each gaze angle bin. This transformation facilitates the estimation of the likelihood of the gaze angle within specific bins. The expected value of the gaze angle, $E(X)$, is then computed by multiplying each bin's index by its corresponding probability and summing these products across all bins, as expressed in the equation:

$$E(X) = \sum_{i=1}^{b} p_i \cdot i \tag{3.4}$$

Here, p_i is the probability of the $i'th$ bin and b is the number of bins. Given that the expected value's range, $[1, b]$, does not directly align with the desired gaze angle ranges for the datasets MPIIFaceGaze, GazeCapture, Gaze360 and RTGene, an adjustment to this formula 3.4 is necessary as follows:

$$\theta_p = w \sum_{i=1}^{b} p_i \left(i - \frac{1+b}{2}\right) \tag{3.5}$$

In this adjusted equation, θ_p represents the continuous gaze angle, and w is the width of each bin, set at 3 degrees in L2CS-Net. This adjustment ensures the mapping of bin indices to the actual angle values, accounting for the datasets' respective angle ranges. By subtracting the index of the bin i from approximately half of the total number of bins ($(1 + b)/a$), you align the bin space with the corresponding angle space.

To refine the estimation of continuous angles, a regression loss function is employed in L2CS-Net. Two common regression losses in gaze estimation are explored including mean squared error MSE and mean absolute error MAE:

$$MSE(\theta_t, \theta_p) = \frac{1}{n} \sum_{i=1}^{n} (\theta_{t_i} - \theta_{p_i})^2 \tag{3.6}$$

$$MAE(\theta_t, \theta_p) = \frac{1}{n} \sum_{i=1}^{n} (\theta_{t_i} - \theta_{p_i}) \tag{3.7}$$

where n is the number of images and θ_{t_i} and θ_{p_i} are the target and predicted angle values for the $i'th$ sample, respectively. The final gaze regression loss L_{reg} using MAE and MSE is calculated as follows:

$$\mathcal{L}_{reg,MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(\theta_{t_i} - w \sum_{j=1}^{b} p_{ij} (j - \frac{1+b}{2}) \right)^2 \tag{3.8}$$

$$\mathcal{L}_{reg,MAE} = \frac{1}{n} \sum_{i=1}^{n} \left(\theta_{t_i} - w \sum_{j=1}^{b} p_{ij} (j - \frac{1+b}{2}) \right) \tag{3.9}$$

3.4.2.3 Regression Classification Loss

The RCS is introduced in L2CS-Net to enhance the estimation of the gaze angle for the pitch and the yaw dimensions. This innovative loss function integrates the benefits of classification and regression losses into a cohesive framework, formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \alpha \cdot \mathcal{L}_{reg} \tag{3.10}$$

In this equation, $\mathcal{L}_{total}$ represents the proposed RCS, with $\mathcal{L}_{cls}$ denoting the classification loss, $\mathcal{L}_{reg}$ representing the regression loss, and α acting as a coefficient that adjusts the contribution of the regression loss to the total loss. Through empirical testing, detailed in Section 3.5, α is set to 1 to achieve optimal gaze estimation performance. This approach underscores the hypothesis to leveraging the complementary strengths of classification and regression to refine gaze prediction accuracy, demonstrating the efficacy of RCS in the experimental evaluations.

The dual-branch architecture and RCS are integrated and optimized jointly, enabling the network to simultaneously learn to classify gaze direction into bins and regress to the exact angles. This integrated approach allows for the effective extraction and utilization of fine-grained features from the eye region, overcoming the challenges associated with the small size of this crucial area in facial images. Through extensive experimentation and validation, L2CS-Net demonstrates superior performance in gaze estimation across various datasets, showing it is effectiveness in capturing the nuances of gaze direction from facial images.

3.5 Experiments and Results

L2CS-Net undergoes extensive evaluations on several benchmark datasets to assess its effectiveness and superiority over existing state-of-the-art methods. In this section, the

Table 3.1: Overview of the four datasets used for assessing L2CS-Net including MPIIFaceGaze, GazeCapture, RTGene and Gaze360. The comparison includes the number of subjects, the gaze range, the number of images, dataset image resolution, whether the dataset is captured outdoors or not, and whether the datasets contain face images or not.

Dataset	**Sub**	**Head Pose**	**Gaze**	**Total**	**Env**
MPIIFaceGaze [252]	15	±15°, ±30°	±20°, ±20°	45k Images	Indoor
GazeCapture [107]	1,474	±30°, ±40°	±20°, ±20°	2.4M Images	Hybrid
RTGene [54]	15	±40°, ±40°	±40°, -40°	123K Images	Indoor
Gaze360 [96]	238	±90°, u/k	±140°, -50°	172K Images	Hybrid

experimental protocols (Section 3.5.1), quantitative results(Section 3.5.2), qualitative results (Section 3.5.3) and comprehensive analysis (Section 3.5.4) are discussed.

3.5.1 Experimental Protocols

This section includes the experimental protocols required to assess the proposed framework including evaluation datasets, performance metrics, data preprocessing, implementation details, and setup of experiments.

3.5.1.1 Evaluation Datasets

The advancement of appearance-based gaze estimation techniques is significantly enhanced by the availability of large-scale datasets, which encompass a wide range of collection methodologies, from laboratory settings to unconstrained environments. The effectiveness of L2CS-Net is validated using four widely used datasets in the gaze estimation domain. These datasets present diverse challenges and settings, facilitating a comprehensive assessment of the model's performance. Table 3.1 provides an overview, detailing key characteristics such as the number of subjects, head pose values, gaze range, dataset size, and the environment. Furthermore, Figure 4.5 presents a visual representation that includes image samples from the four gaze datasets with a diversity of environmental conditions, illuminations, head pose and gaze direction. Through the use of these diverse datasets, L2CS-Net is subjected to comprehensive testing, ensuring its adaptability and performance under various conditions and scenarios.

- **MPIIFaceGaze** [248]: encompasses 213,659 images from 15 participants, capturing a wide array of real-life scenarios, including varying illuminations and environments, through everyday laptop use over three months. This dataset stands out by offering a level of diversity in appearance, head poses, and gaze directions previously unseen in data collected under controlled laboratory conditions. By

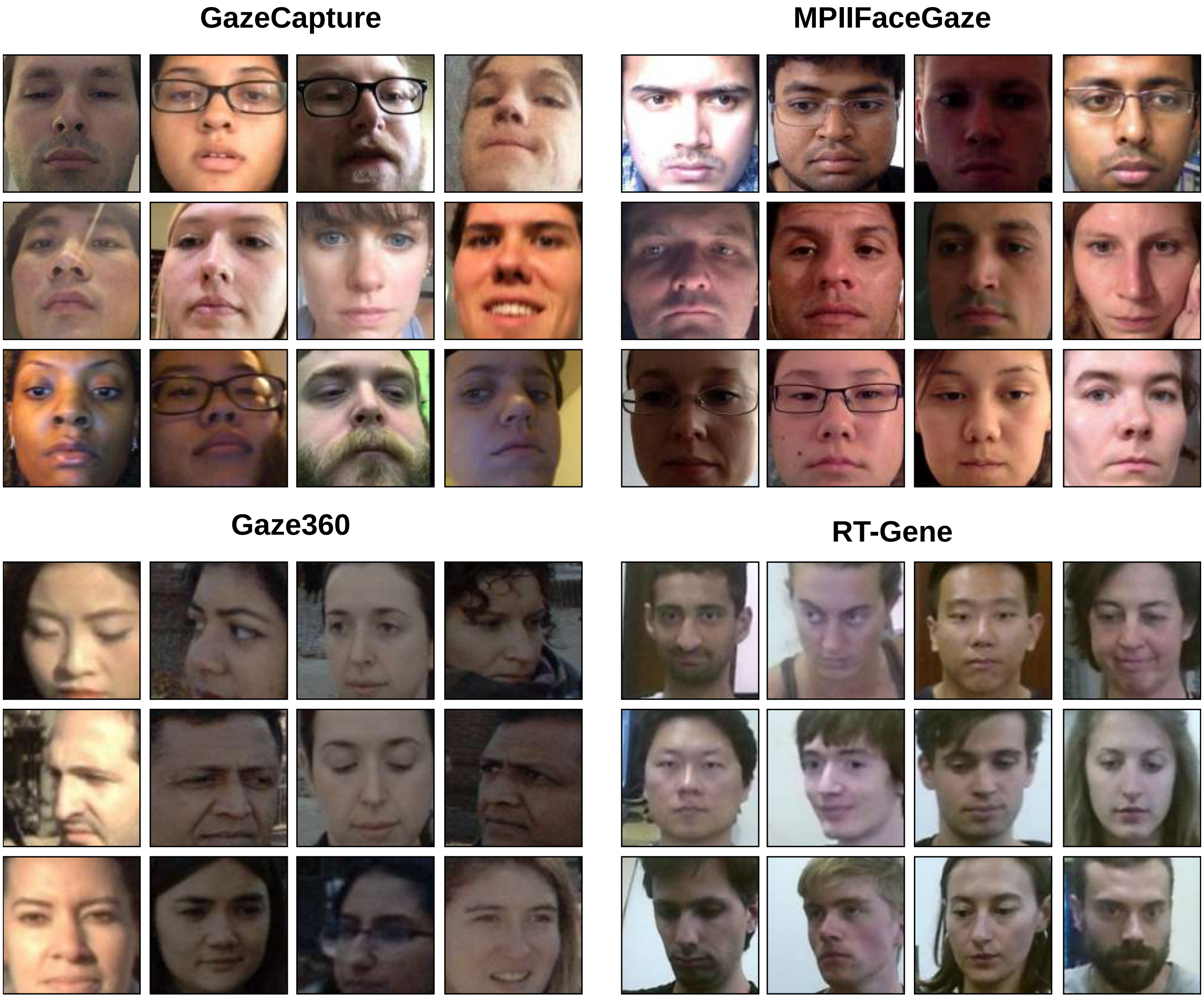


Figure 3.3: Face image samples from the four gaze datasets with diversity of environmental conditions, illuminations, head pose and gaze direction.

providing a rich tapestry of real-world conditions, the MPIIFaceGaze dataset fills a critical gap in gaze estimation research. It offer a comprehensive resource that significantly enhances the ability to develop and test gaze estimation models in scenarios that closely mimic everyday settings.

- **GazeCapture** [107]: marks a significant step forward in the domain of eye tracking, enabling eye tracking technologies on commonly used devices such as smartphones and tablets without the need for additional sensors. It is the first dataset of its kind, comprising nearly 2.5 million frames collected from over 1450 participants, showcasing an unparalleled level of diversity and volume. This vast collection captures a wide array of participant backgrounds, lighting conditions, and head movements, greatly enriching the data available for enhancing the accuracy and generalizability of eye tracking models. The scale and variety offered by the GazeCapture dataset represent a pivotal resource for advancing eye tracking technologies, providing extensive data for research and development efforts in a multitude of applications.

- **Gaze360** [95]: stands out as the largest of its kind for 3D gaze estimation with 172K images from 238 subjects of different genders, ages, and ethnicity. It covers an extensive range of head poses and distances in a variety of indoor and outdoor settings, made possible through an efficient data collection methodology. This diversity and scale enable the dataset to significantly advance the study and application of gaze estimation in unconstrained environments. The images were captured in various indoor and outdoor settings using a Ladybug multi-camera system, presenting a substantial test for unconstrained gaze estimation models. By providing such a comprehensive dataset, researchers address the critical need for large and diverse annotated training data, which has been a limiting factor in the field. The Gaze360 dataset not only facilitates the development of more accurate and generalizable gaze estimation models, but also supports cross-dataset evaluations and domain adaptation research, marking a notable contribution to gaze estimation research and its practical applications.

- **RTGene** [54]: consists of 122,531 labeled images and 154,755 unlabeled images of 15 participants. It addresses the challenge of accurately estimating gaze over large distances and over a broad range of head poses and eye gaze angles. This dataset features a comprehensive collection of images with ground-truth annotations for gaze and head poses. These annotations obtained using a combination of a motion capture system for head pose and mobile eye-tracking glasses for eye gaze. A novel aspect of this dataset is the semantic in-painting technique used to replace the appearance of the eye-tracking glasses with realistic skin texture in the images to ensure the dataset's relevance for both training and testing.

3.5.1.2 Performance Metrics

The primary metric for assessing the effectiveness of gaze estimation models is the angular error between the estimated gaze direction and the ground truth in all test images. This error quantifies the accuracy of the model, with a smaller angle error indicating a more precise estimate. The model outputs yaw and pitch angles, denoted θ (pitch) and ϕ (yaw), which are then converted into a 3D gaze direction vector $g = (g_x, g_y, g_z)$ using the following transformations:

$$g_x = -\cos(\phi) \cdot \sin(\theta) \tag{3.11}$$

$$g_y = -\sin(\phi) \tag{3.12}$$

$$g_x = -\cos(\phi) \cdot \cos(\theta) \tag{3.13}$$

Given the ground truth gaze direction g and the predicted gaze direction ĝ, the angular error $\mathcal{L}_{angular}$ can be calculated as:

$$\mathcal{L}_{angular} = \arccos\left(\frac{\mathrm{g} \cdot \hat{\mathrm{g}}}{\|\mathrm{g}\|\|\hat{\mathrm{g}}\|}\right) \tag{3.14}$$

In addition to accuracy, the computational complexity of gaze estimation models is a critical metric, especially in real-time application scenarios. To objectively measure this aspect, experiments are conducted across different models using the hardware setups. This approach ensures a fair comparison of computational efficiency, highlighting the balance each model strikes between accuracy and speed.

3.5.1.3 Data Preprocessing

To prepare the dataset images for L2CS-Net model, normalization procedures are executed as outlined in previous research [245]. This process involves adjusting the virtual camera's position through rotation and translation to negate the roll angle of the head and maintain a consistent distance between the camera and the center of the face. This step is crucial for aligning the datasets with the input requirements of L2CS-Net. For Gaze360, RTGene and MPIIFaceGaze, the same procedures as in [34, 245] are followed to preprocess the dataset and create normalized face crops. For GazeCapture, the settings described in [231, 245] are employed to create normalized head crops.

To integrate the proposed RCS function effectively, all the datasets images are relabeled to reflect new bin classes. This involves converting the continuous gaze angles (both pitch and yaw) into discrete bin labels using one-hot encoding, tailored to the specific gaze range annotations of each dataset. For the Gaze360 dataset, 90 bins are established, each spanning 3 degrees, to encompass its wide gaze range of [-140°, 140°]. Similarly, RTGene, MPIIFaceGaze, and GazeCapture are segmented into 30, 14 and 14 bins, respectively, based on the same 3-degree width, covering their gaze ranges of [-40°, 40°] for RTGene, [-20°, 20°] for MPIIFaceGaze and [-20°, 20°] for GazeCapture. This binning process generates two types of target annotation for the datasets: continuous and binned labels, making them compatible with the model's requirement for the multi-loss approach of regression and classification losses. This binning strategy ensures that the model can be trained and evaluated on a standardized and comprehensive representation of gaze data, facilitating robust learning across diverse gaze values.

3.5.1.4 Implementation Details

The experiments are conducted using the PyTorch framework (version 1.8.1) on a state-of-the-art computing setup equipped with an Intel(R) Core (TM) i7-7800X CPU, an NVIDIA RTX 3080 with 12GB of memory. For training L2CS-Net, the ImageNet-1K [159] dataset is utilized for pretraining L2CS-Net to benefit from a wide and diverse visual representation. The EfficientNetV2-S model serves as the backbone of the network architecture, as it has a balance between efficiency and performance. The network processes facial images with a resolution of 224×224. The network undergoes training for a total of 50 epochs, with a batch size of 16.The Adam optimizer is used with a learning rate of $1e^{-5}$. The evaluation of network performance is conducted rigorously on the MPIIFaceGaze, GazeCapture, Gaze360, and RTGene datasets. These evaluations adhere to the specific methodologies outlined in Sec.3.5.1.1, ensuring a fair assessment of the network's gaze estimation capabilities compared with state-of-the-art methods.

3.5.1.5 Implementation Details

The experiments are conducted using the PyTorch framework (version 1.8.1) on a state-of-the-art computing setup equipped with an Intel(R) Core (TM) i7-7800X CPU, an NVIDIA RTX 3080 and 32GB of RAM. For training the network, the ImageNet-1K [159] dataset is utilized for pretraining L2CS-Net to benefit from a wide and diverse visual representation. The EfficientNetV2-S model serves as the backbone of the network architecture, as it has a balance between efficiency and performance. The network processes facial images with a resolution of 224×224. The network undergoes training for a total of 50 epochs, with a batch size of 16. The Adam optimizer is used with a learning rate of $1e^{-5}$. The evaluation of network performance is conducted rigorously on the MPIIFaceGaze, GazeCapture, Gaze360, and RTGene datasets. These evaluations adhere to the specific methodologies outlined in Sec.3.5.1.1, ensuring a fair assessment of the network's gaze estimation capabilities compared with state-of-the-art methods.

3.5.1.6 Experiments Setup

To evaluate the accuracy of L2CS-Net, a detailed experiments are conducted that focuses on within dataset evaluation in the four mentioned datasets. Each dataset had different evaluation protocols based on previous research to ensure that the performance measures are reliable and comparable. Consequently, four within dataset evaluation tasks are established include:

- $\boldsymbol{\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}}$: In this task, 15-fold cross-validation is used to evaluate L2CS-Net in the MPIIFaceGaze dataset. The model is trained on data from all 15 subjects except one, which is held as the test set. This procedure is repeated in such a way that each subject is used as a test subject once, ensuring that the evaluation covers the variability between different individuals in a comprehensive way.

- $\boldsymbol{\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}}$: In this task, a predefined split of 1379083 image training set, 191842 image testing set, and 63518 image validation set as specified by state-of-the-art. This method ensures that the model is tested against a wide range of real-world conditions since GazeCapture includes a variety of indoor and outdoor scenes collected from thousands of different subjects.

- $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}$: In this task, a predefined split of 84000 training, 16500 testing, and 500 validation data as specified by the dataset authors are used. This split ensures that the model is evaluated against a diverse set of subjects and environmental conditions, reflecting its robustness and adaptability in varied scenarios.
- $\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}$: In this task, 3-fold cross-validation is used to evaluate L2CS-Net on RTGene dataset. The subjects are divided into three groups and each group is used alternately as a test set while the remaining subjects are used for training. This setup tests the performance of the model across different subsets of data, highlighting its effectiveness in real world settings where user variability is significant.

These four tasks are chosen based on different datasets with diversity in capture conditions, subject variability, and range of gaze angles, making them ideal for assessing the robustness of L2CS-Net. Notably, the MPIIFaceGaze and GazeCapture datasets have a relatively narrow gaze range, extending approximately [-20°, 20°] degrees. In contrast, the RTGene and Gaze360 datasets encompass a wider range, with the gaze range extending to [-40°, 40°] and [-140°, 140°] degrees, respectively.

3.5.2 Quantitative Results

This section presents the quantitative results of the experiments, demonstrating the superiority of L2CS-Net in improving the accuracy of gaze estimation over existing state-of-the-art techniques in within dataset evaluation.

3.5.2.1 Within Dataset Evaluation

The proposed L2CS-Net is evaluated on the four within dataset evaluation tasks including $\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}$, $\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}$, $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}$ and $\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}$. As mentioned in Section 3.5.1.6, each task is evaluated based on specific criteria to provide a fair comparison with state-of-the-art methods. The accuracy of L2CS-Net is evaluated against several state-of-the-art gaze estimation methods. The comparison includes a variety of approaches that utilize both face and eye images, including DilatedNet [27], Full-Face [251], CA-Net [30], RTGene [54], AGE-Net [14], Fare-Net [36] and L2DNet [133]. These methods demonstrate the traditional approach of integrating information from face and eye to predict gaze direction. In contrast, the comparison also includes methods that rely solely on facial images, such as Gaze360 [95], ETH-Gaze [242], and SAtten-Net [135]. This category underscores the evolving trend towards models that aim for efficiency without significantly compromising accuracy. Moreover, the comparison includes the diversity in network architectures employed in these methods, from conventional CNNs and dilated CNNs to advanced architectures such as transformers with self-attention mechanisms. This comparative analysis is critical for showing the advancements of L2CS-Net in terms of methodology and quantitative performance

Table 3.2: Comparison with the state-of-the-art methods in within dataset evaluation: L2CS-Net achieves state-of-the-art gaze performance in the four datasets evaluation tasks with mean angular errors of 3.86°, 2.71°, 10.12°, and 6.50° on MPIIFaceGaze, GazeCapture, Gaze360, and RTGene datasets, respectively.

Methods	$\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}$ (15-fold)	$\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}$ (test-set)	$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}$ (test-set)	$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}$ (3-fold)
FullFace [251]	4.90°	-	-	7.44°
AGE-Net [14]	4.09°	-	-	7.44°
GazeTR-Pure [32]	4.74°	-	13.58°	8.06°
L2DNet [133]	4.30°	-	-	8.40°
Fare-Net [36]	4.30°	-	-	8.40°
SAtten-Net [135]	**4.04°**	-	**10.70°**	**7.00°**
CA-Net [30]	4.27°	-	11.20°	8.27°
RTGene [54]	4.30°	-	12.26°	8.00°
MANNet [215]	4.30°	-	-	13.20°
ETH-Gaze [242]	4.80°	**3.30°**	-	12.00°
Gaze360 [95]	**4.06°**	-	**11.20°**	**7.12°**
FAZE [144]	-	3.49°	-	-
RSN [247]	4.50°	**3.32°**	-	-
GEDDNet [28]	4.69°	-	-	8.17°
Dilated-Net [27]	4.42°	-	13.73°	8.38°
EM-Gaze [215]	4.10°	-	-	-
L2CS-Net	**3.86°**	**2.71°**	**10.12°**	**6.50°**

across standard datasets. For a fair comparison, the results of all compared methods are either from the original papers or obtained by running their open-source codes on the same hardware settings. Table 3.2 presents a comparative analysis of mean angular error, comparing L2CS-Net with existing state-of-the-art gaze estimation methods on the four within dataset tasks. The top three best gaze estimation performance results are highlighted in bold.

For $\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}$ task, previous methods struggled to reduce the error rate below the 4.04° reported by SAtten-Net [135]. This model integrates a self-attention module and outperforms traditional CNN-based approaches by a significant margin due to its ability to effectively learn data with high variance. However, L2CS-Net has successfully reduced the mean angular error to 3.86°, outperforming SAtten-Net by approximately 5%. Furthermore, L2CS-Net surpass the 4.06° result reported by Gaze360 [95] which utilize 7-frames LSTM network for predicting gaze values. Since the GazeCapture dataset only provides gaze labels on a 2D screen, many methods have ignored it in their evaluation. A crucial preprocessing step, performed as detailed in [231], converts these 2D labels into 3D gaze labels using head pose values. L2CS-Net follows this preprocessing step to conduct experiments on the GazeCapture dataset for the $\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}$ task. L2CS-Net achieves a mean angular error of 2.71°, surpassing the existing state-of-the-art result of 3.30° and 3.32° by ETH-Gaze [242] and RSN [247], marking a 18% improvement and establishing L2CS-Net as the new benchmark in the $\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}$ task.

For the $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}$ task, L2CS-Net has notably surpasses existing models and achieve state of the art performance with 10.12° mean angular error. In particular, it outperforms the best result previously reported by the SAtten-Net model 6%, which utilizes the self-attention mechanism. For the $\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}$ task, L2CS-Net has successfully achieved state-of-the-art gaze performance with a mean angular error of 6.50°. Notably, it outperforms the previously best reported result of 7.00° and 7.12° by SAtten-Net and Gaze360, marking up to 8% improvement in gaze estimation accuracy.

In conclusion, while methods incorporating advanced architectural design, such as attention mechanisms (e.g., SAtten-Net [135]) and temporal data (e.g., Gaze360 [95]) demonstrate promising results in various datasets. However, their performance is often mitigated by higher computational demands and larger model sizes, which may restrict their applicability in resource-constrained environments. The proposed L2CS-Net, achieve state-of-the-art performance on the four within dataset evaluation tasks with up to 18% improvement in gaze accuracy, illustrating its effectiveness in leveraging fine-grained gaze features to surpass state-of-the-art models. This proves that L2CS-Net is able to accurately extract fine-grained features from the eye appearance.

3.5.2.2 Computational Efficiency

To evaluate the computational efficiency of L2CS-Net, it is compared against various state-of-the-art gaze estimation methods focusing on processing time, parameter count, and floating-point operations per second (FLOPs). As shown in Table 3.3, methods are categorized based on model complexity into high-parameter models with over 30 million parameters, mid-range models with 5 to 30 million parameters, and compact models with fewer than 5 million parameters. L2CS-Net achieve the lowest FLOPs and processing time among the compared methods. Notably, while it had a larger parameter count than the Gaze360 [95] approach, L2CS-Net significantly reduced computational costs by approximately 77% fewer FLOPs and improved processing time by 4% over Gaze360 [95]. By optimizing the network architecture and leveraging a face-only input stream, L2CS-Net has significantly enhanced both the accuracy and speed of gaze estimation, setting a new benchmark in gaze estimation accuracy and efficiency.

Table 3.3: Comparison with the state-of-the-art-methods methods in terms of computational complexity: L2CS-Net achieve the lowest FLOPs and processing time among the other methods while having large number of parameters.

Methods	**Backbone**	**#Params**	**#FLOPs**	**Time**
		[M]	[G]	[ms]
FullFace [251]	CNN	196.6	2.99	-
AGE-Net [14]	Tr	109.0	35.75	66
GazeTR-Pure [32]	Tr	104.0	58.3	-
L2DNet [133]	Dilated-CNN	87.0	-	-
Fare-Net [36]	CNN	75.0	27.5	49
SAtten-Net [135]	Tr	74.8	19.7	379
CA-Net [30]	CNN	34.0	15.6	30
RTGene [54]	CNN	31.0	30.81	35
MANNet [215]	CNN	29.5	2.7	-
ETH-Gaze [242]	CNN	23.8	4.12	6.4
Gaze360 [95]	RNN	14.6	12.78	5.4
CDBN [270]	CNN	13.0	-	8.4
FAZE [144]	CNN	-	-	-
RSN [247]	CNN	-	-	-
GEDDNet [28]	Dilated-CNN	4.0	-	6.3
Dilated-Net [27]	Dilated-CNN	3.9	3.1	6.7
EM-Gaze [215]	CNN	2.7	-	7.3
L2CS-Net	**CNN**	**22.4**	**2.97**	**5.2**

3.5.2.3 Ablation Study

The comprehensive ablation study aims to validate the individual contributions of different components within the gaze estimation model, specifically focusing on loss functions, network architecture, and the choice of backbone. Through systematic experiments on $\boldsymbol{\mathcal{D}_{MP}} \rightarrow \boldsymbol{\mathcal{D}_{MP}}$, $\boldsymbol{\mathcal{D}_{GC}} \rightarrow \boldsymbol{\mathcal{D}_{GC}}$, $\boldsymbol{\mathcal{D}_{GZ}} \rightarrow \boldsymbol{\mathcal{D}_{GZ}}$ and $\boldsymbol{\mathcal{D}_{RT}} \rightarrow \boldsymbol{\mathcal{D}_{RT}}$, the impact of these elements on overall performance is evaluated.

1) Contribution of the RCS Function: To effectively evaluate the effectiveness of the proposed RCS, an experiment is conducted to compare five different networks each utilizing a different loss function. Each network uses the same backbone architecture as EfficientNetV2 to ensure that performance differences are only due to the loss function used rather than architectural differences. Two basic networks are built for this comparison one uses the MAE as the loss function, and the second uses the

Table 3.4: Ablation study on losses: the proposed RCS function improves the gaze performance of L2CS-Net compared with conventional MSE, MAE and CE losses.

Networks	$\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}$	$\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}$	$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}$	$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}$
CE-Loss	4.98°	3.48	11.70°	8.55°
MSE-Loss	4.80°	3.07°	11.30°	8.12°
RCS (MSE)	**3.92°**	**2.80**	**10.23°**	**6.57°**
MAE-Loss	4.75°	2.98°	11.52°	7.76°
RCS (MAE)	**3.86°**	**2.71°**	**10.12°**	**6.50°**

MSE since they are standard choices in regression tasks for their simplicity and effectiveness. To explore the effect of classification loss on gaze estimation, a third network is introduced that makes use of cross entropy loss (CE) solely, omitting any regression component. This network aims to understand the extent to which classification alone contributes to gaze estimation.

Against these baselines, the proposed RCS that combines regression and classification losses is utilized. For a fair comparison, two networks are proposed based on the RCS one with MAE regression component and the other utilizing the MSE part. This approach ensures that each comparison is fair and directly evaluates the effect of combining regression and classification in the loss function. The experiments revealed a significant improvement in mean angular error when employing RCS, demonstrating its superior capability in enhancing gaze estimation accuracy (Table 3.4). Notably, integrating MAE within the RCS framework decreases the mean angular error by up to 19% compared to traditional loss functions, underlining the value of combining regression and classification components. The results demonstrate that the proposed RCS significantly outperforms traditional loss functions in gaze estimation tasks. Various hyperparameters are employed to identify the optimal settings for maximizing gaze estimation performance across the networks. For example, the best gaze performance for conventional networks, which utilize MAE and MSE as loss functions, is achieved with a learning rate of $1e^{-4}$. In contrast, for the network employing the RCS, a lower learning rate of $1e^{-5}$ results in the highest performance.

2) Contribution of the Network Architecture: To robustly evaluate the effectiveness of the proposed L2CS-Net architecture, two different networks are proposed using the same backbone architecture as EfficientNetV2 and the same loss function, RCS. The first network, named Conventional-1F, uses only one fully connected layer for gaze angle prediction pitch and yaw. The second network, the proposed L2CS-Net, incorporates two separate fully connected layers for separate gaze angle prediction. Results are presented in Table 3.5, which compares the performance of the two networks. The results from this comparison clearly demonstrate superior gaze estimation performance of the proposed L2CS-Net, evidencing the advantages of separating the

Table 3.5: Ablation study on network architecture. The proposed two branch CNN architecture improves the gaze performance compared with Conventional-1F network.

Networks	$\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}$	$\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}$	$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}$	$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}$
Conventional-1F	4.02°	3.12	10.53°	6.95°
Proposed	**3.86°**	**2.71°**	**10.12°**	**6.50°**

Table 3.6: Ablation study on backbones. Comparison of the angular error between the different backbones using L2CS-Net.

Networks	$\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}$	$\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}$	$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}$	$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}$
ResNet18	4.00°	3.15°	10.7°	6.82°
ResNet50	3.92°	2.90°	10.23°	6.67°
EfficientNet	**3.86°**	**2.71°**	**10.12°**	**6.50°**

prediction of each gaze angle with a dedicated fully connected layer. This improvement in performance can be attributed to the model's ability to independently capture and learn the fine-grained features specific to each gaze angle, pitch, and yaw, thereby achieving a more accurate and refined estimation. In summary, the separation of the pitch and yaw branches with two fully connected layers decreases the mean angular error up to 14% compared with Conventional-1F.

3) Contribution of the Backbone: To effectively assess the EfficientNet backbone in L2CS-Net, two backbones widely used in previous research: ResNet-18 and ResNet-50 are adapted. In Table 3.6, the results of L2CS-Net with ResNet-18, ResNet-50 and EfficientNetV2 are presented. In particular, L2CS-Net even when utilize the approximately 50% smaller ResNet-18 backbone, still outperforms all other methods listed in Table 3.2 across the four within dataset tasks. These results confirmed that the exceptional performance of L2CS-Net is primarily come from the network architecture and loss function, rather than the specific backbone utilized. Additionally, the results with ResNet-50 show an improvement in gaze estimation accuracy over those reported by ETH-Net[242], which employs the same backbone. Finally, it becomes apparent that L2CS-Net with EfficientNetV2 surpasses the ResNet-50 model in terms of both performance and computational efficiency, despite having a comparable number of parameters (as detailed in Table 3.2). This underlines the effectiveness and efficiency of L2CS-Net method, validating its robustness across different backbone architectures.

Overall, the ablation study underscores the critical role of the RCS function, the dual fully connected layer architecture, and the EfficientNet backbone in pushing the limits of gaze estimation accuracy.

3.5.3 Qualitative Results

While the quantitative results establish the superior performance of L2CS-Net in numerical terms, the qualitative analysis provides invaluable insights into the model's practical capabilities and the nuances of its predictions. This section delves into the visual and interpretative aspects of the performance of L2CS-Net, illustrating its effectiveness through various examples and visualizations.

3.5.3.1 Visualization of Gaze Predictions

A series of visualizations is presented showing the gaze predictions of L2CS-Net across a range of challenging scenarios, including variations in lighting, subject head pose, and background complexity. These examples demonstrate the robustness of L2CS-Net and its ability to accurately estimate gaze direction even in the presence of potential challenges. Side-by-side comparisons of the model's predictions with the ground truth data highlight its accuracy. The visual alignment between the predicted gaze vectors and the actual gaze directions of the subjects underscores the model's accuracy in capturing the rich gaze information.

3.5.3.2 Visualization of Gaze features.

Class Activation Maps (CAMs)[259] are utilized in both conventional gaze estimation methods and the proposed L2CS-Net to visualize the model's attention areas during the gaze estimation process. The conventional method employs a single fully connected layer alongside MAE for regression loss, whereas L2CS-Net integrates dual fully connected layers and adopts a multi-loss approach. Illustrated in Figure 3.5, these visualizations provide deep insight into the operational nuances of L2CS-Net.

A critical observation from the CAM visualizations is the visible focus of L2CS-Net on crucial eye regions while simultaneously leveraging global features to predict gaze angles. This ability demonstrates the refined capacity of L2CS-Net to extract and utilize the most informative features for gaze estimation. This observation confirms the importance of the dual-branch architecture and the proposed multi-loss function. The adoption of a multi-loss approach in L2CS-Net allows for more targeted and refined feature extraction, as evidenced by CAMs that display a more focused attention pattern on relevant eye region features. This strategy notably emphasizes the importance of eye regions within visual input, consistent with the intuitive understanding that these areas contain crucial cues for accurate gaze estimation. By prioritizing these regions, L2CS-Net demonstrates a superior ability to capture the essence of gaze direction, reinforcing the significance of these methodological choices in advancing gaze estimation accuracy.

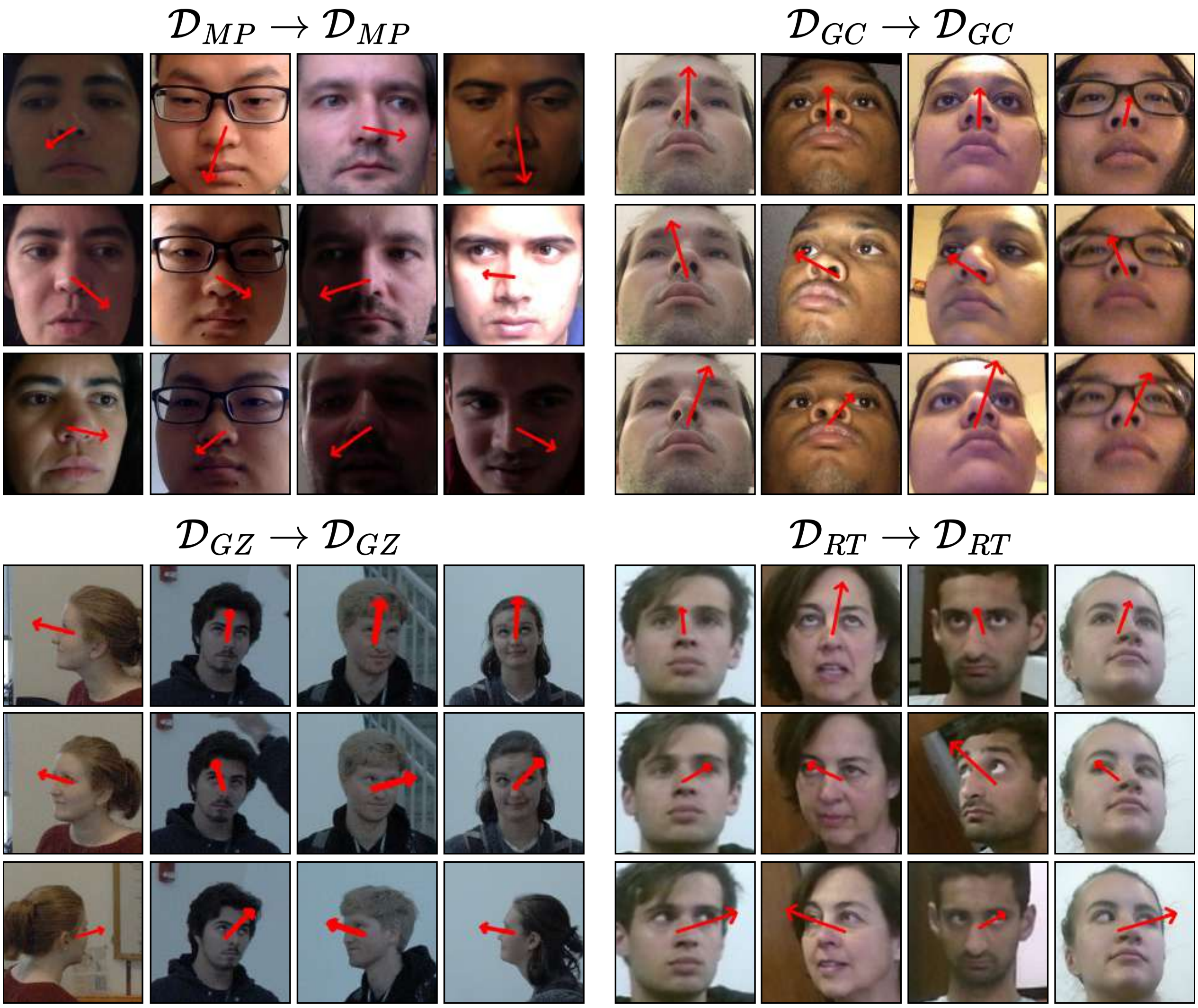

$\mathcal{D}_{MP} \to \mathcal{D}_{MP}$
$\mathcal{D}_{GC} \to \mathcal{D}_{GC}$
$\mathcal{D}_{GZ} \to \mathcal{D}_{GZ}$
$\mathcal{D}_{RT} \to \mathcal{D}_{RT}$

Figure 3.4: Visualization of Gaze Predictions of L2CS-Net on face images on the four within dataset tasks.

3.5.4 Comprehensive Analysis

This section delves into the performance of L2CS-Net, offering a deeper understanding of its capabilities through subject-wise performance analysis and the strategic use of classification bins.

3.5.4.1 Subject-Wise Performance Analysis on MPIIFaceGaze

In a further extension of L2CS-Net evaluation, the focus shifts from aggregate performance measures to a detailed examination of the robustness of the model across individual subjects within the MPIIFaceGaze dataset. This dataset, which includes data from 15 distinct subjects, offers a unique opportunity to evaluate the effectiveness of gaze estimation methods on a subject-by-subject basis, thus providing insights into their adaptability and accuracy across diverse environmental conditions.

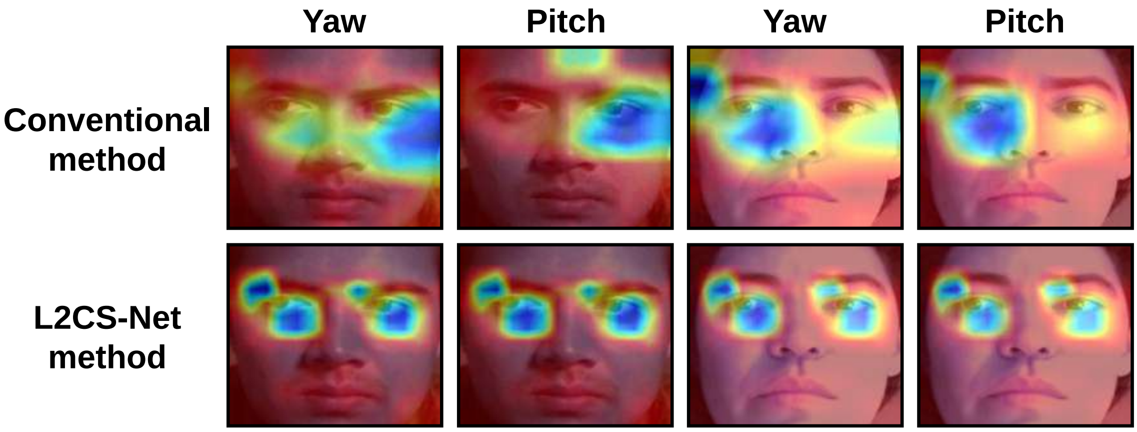


Figure 3.5: CAM visualization displays two subjects from the MPIIFaceGaze dataset. The first row presents results from a conventional gaze estimation method using MSE regression loss only, while the second row features results from the dual-branch L2CS-Net employing a multi-loss approach. The images depict yaw and pitch, respectively, in order from left to right.

The analysis delves into the subject gaze accuracy achieved by the proposed L2CS-Net, comparing these results with those obtained by FARE-Net[36], a previous model that also reports performance at the individual subject level. The comparison results, visually summarized in Figure 3.6, reveal that the proposed method outperforms FARE-Net in terms of gaze accuracy for 11 out of 15 subjects and provide comparable results in the remaining four subjects.

This relative advantage not only underscores the superior accuracy of the proposed method but also emphasizes its remarkable robustness. The ability of the model to consistently outperform established benchmarks in the majority of subjects demonstrates its effectiveness in accommodating inherent variability between individuals. Whether these differences are due to different eye shapes, facial features, or differences in interaction environments, L2CS-Net shows a remarkable ability to adapt and maintain high levels of accuracy. These results are important, as they confirm the robustness and reliability of the model across a wide range of users, representing a major step forward in the pursuit of effective gaze estimation solutions.

3.5.4.2 Impact of Classification Bins

The implementation of classification bins within the multi-loss function plays a pivotal role in achieving fine-grained gaze estimation accuracy. This section explores the impact of these classification bins on L2CS-Net performance.

L2CS-Net can initially narrow down the gaze direction to a specific angular interval (coarse estimation) before refining this prediction to achieve a high level of accuracy (fine estimation). This two-step approach significantly enhances the accuracy of the model. The choice of bin size and range is critical to balancing the granularity of gaze estimation with the model's ability to generalize across various gaze values. To

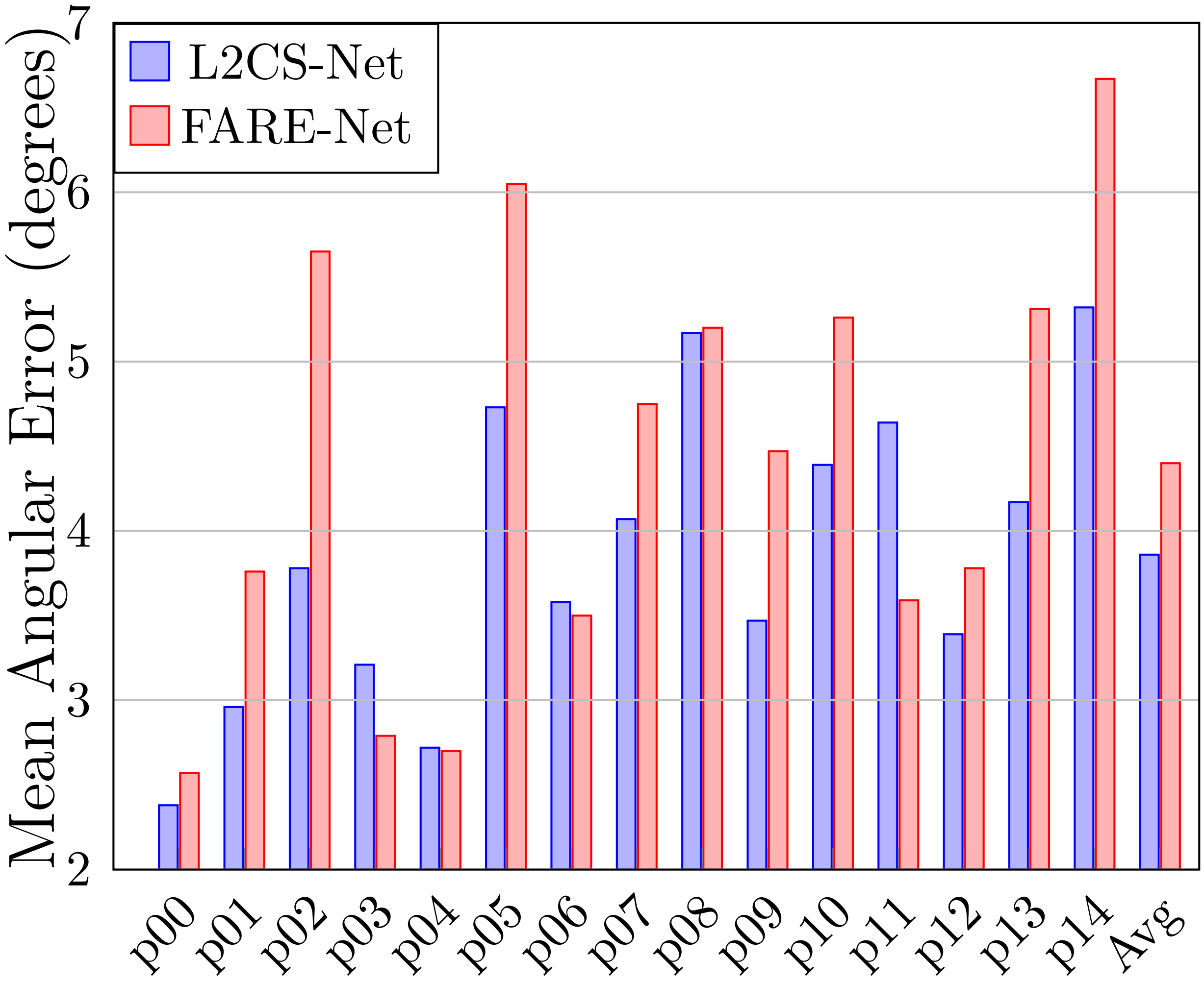


Figure 3.6: Comparison of subject-wise gaze accuracy between L2CS-Net and FARE-Net [36] on MPIIFaceGaze dataset.

adjust the best bin size, extensive experiments on the four within datasets tasks are performed with chaining the bin width from 0 to 10. The results of these experiments are visualized in Figure.3.7. The figure indicate that a bin width of 3 degrees offers an optimal balance, enabling the model to accurately categorize and refine gaze estimations without overwhelming the system with too many categories or losing accuracy due to overly broad intervals. The flexibility to adjust the number and range of bins based on the specific dataset and its inherent gaze angle variability is a crucial aspect of model design. This dynamic adjustment ensures that the framework remains effective across different datasets, each with its unique challenges and requirements.

3.6 Conclusion

This chapter outlines the development of L2CS-Net, a two-branch, multi-loss CNN approach specifically designed for the direct estimation of gaze direction from facial

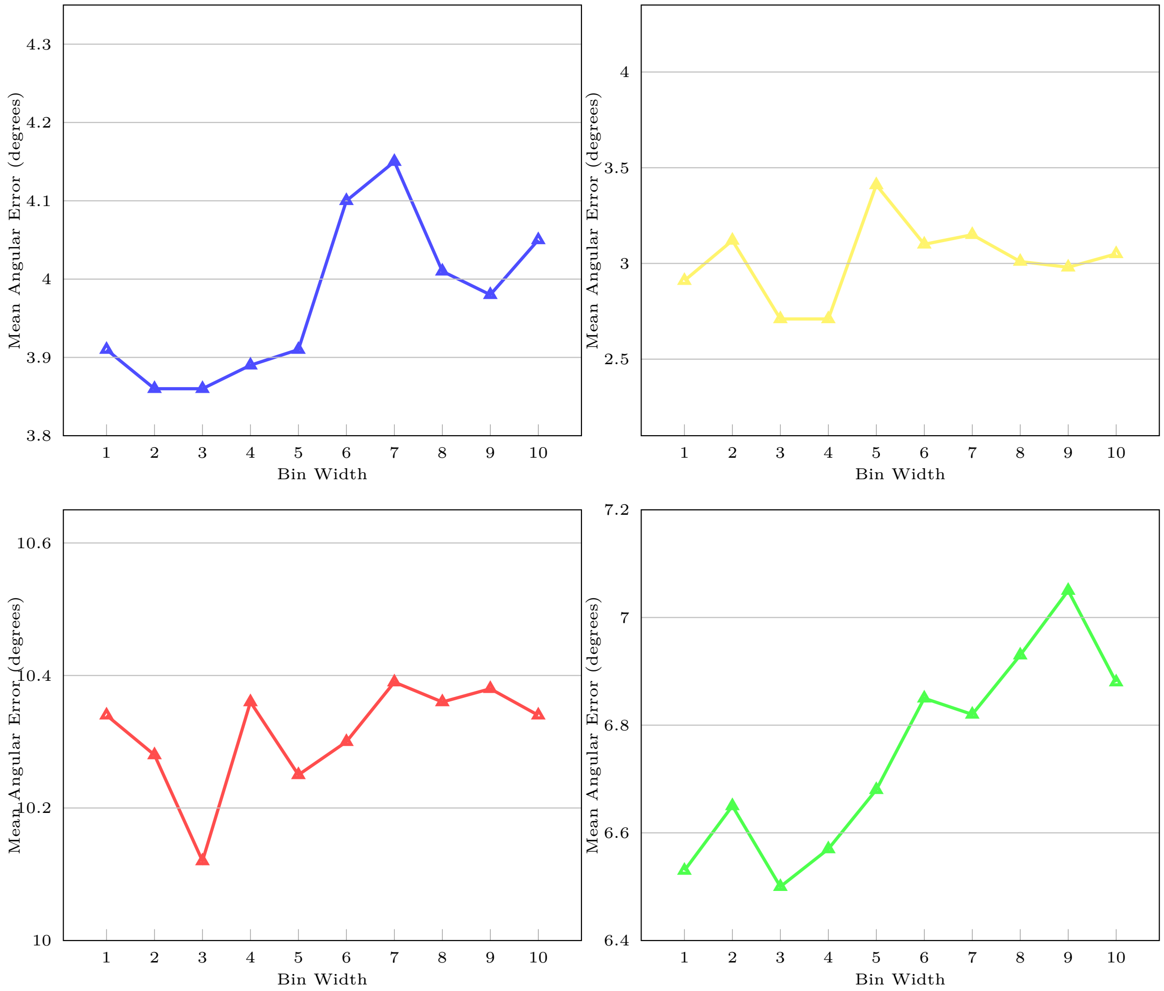


Figure 3.7: Impact of changing the bin width on the gaze performance for the four within dataset tasks datasets. (a) MPIIFaceGaze dataset, (b) GazeCapture dataset, (c) Gaze360 dataset, and (d) RTGene dataset.

images. The model uniquely predicts each gaze angle pitch and yaw using dedicated fully connected layers. This strategy is vital for isolating and enhancing the nuanced features specific to each gaze angle, thereby improving the accuracy of the model.

To enhance the accuracy of gaze direction predictions, L2CS-Net employs a novel multi-loss function for each gaze angle. This function merges regression and classification losses to enable a joint optimization process that significantly boosts the feature extraction process. As a result, L2CS-Net can extract more informative gaze features, greatly enhancing the overall performance of gaze estimation. L2CS-Net employs a softmax layer along with cross-entropy loss to facilitate coarse gaze predictions. This setup is refined further in the gaze regression phase, where the expectation of gaze class probabilities is calculated and refined by a gaze regression loss, ensuring accurate gaze estimations.

The efficacy and robustness of the proposed L2CS-Net have been rigorously validated across four of the most challenging and unconstrained gaze datasets currently available: MPIIFaceGaze, GazeCapture, Gaze360, and RTGene. Through this extensive validation process, L2CS-Net achieves state of art accuracy in gaze estimation demonstrating up to a 18% improvement compared with the reported state of the art benchmark.

4 Improving Gaze Estimation Reliability Using Multi-task Learning

4.1 Introduction

Single image face analysis represents a dynamic area within computer vision, drawing considerable interest for its critical applications in various fields. Estimating human gaze from a single image is essential to advance human-robot interaction [2, 15, 169], autonomous driving [59, 186] and virtual reality[39, 149, 150]. These applications leverage gaze estimation models to interpret human intention accurately, facilitating seamless interaction between humans and machines or virtual environments.

A fundamental challenge toward robust gaze estimation is achieving high accuracy across different domains [29, 135, 220, 221]. The diversity of human subjects, environmental settings, and data acquisition methodologies in existing datasets presents a significant challenge. Datasets such as MPIIFaceGaze, GazeCapture, and RT-GENE, have been crucial in advancing gaze research but also highlight the variability in gaze values across different demographic groups and settings. The variance in camera specifications, resolutions, and lighting conditions further increase the challenge, introducing biases that compromise the ability of models to generalize. Moreover, the process of annotating gaze data with precise 3D labels is fraught[86, 95, 248] with potential inaccuracies and introduces noise, which can degrade model performance. This scenario underscores the requirement for models that can navigate these diversities and maintain high performance in unseen environments.

Another challenge in robust gaze estimation is the intricate relationship between eye and head movements[35, 58]. The gaze direction of a person is influenced not only by the movement of their eyes but also by the orientation of their head. This dynamic interplay requires a modeling approach that can accurately capture the combined effect of eye and head movements on gaze direction. Previous studies have approached this challenge either by implicitly learning the relationship from extensive data or by explicitly incorporating head pose information into gaze estimation models.

This chapter proposes a different strategy to address the challenge of improving the reliability of gaze performance across different domains. Two pivotal strategies that have been used for this issue include domain generalization and domain adaptation. These approaches aim to enhance the robustness of gaze estimation models by enabling them to maintain high accuracy across diverse datasets. Domain generalization techniques [29, 135] aim to learn domain-invariant features by using data exclusively from the source domains, without access to data from the target domain. However, their performance still prevails need for improvement due to a lack of diverse, annotated real-world datasets. In contrast, domain adaptation [11, 145, 166, 198, 231] uses data from the target domain to minimize the disparity between the source domains and the target domain. However, these methods depend on extra models or annotations, potentially limiting their practicality in real-world settings. Moreover, while these methods are effective for specific domains, the resulting models lack generalizability to entirely unseen domains, as they are fine-tuned for particular target environments.

In this chapter, MTGH-Net is presented as an innovative solution to the reliability of gaze estimation. MTGH-Net adopts a multi-task learning framework designed to leverage the inherent correlation between gaze and head pose estimation tasks. This approach aims to enhance the robustness and generalization capabilities of gaze estimation models across various domains. The gaze and head pose tasks each have their own datasets characterized by diverse labels and environmental factors, yet there is no existing large-scale dataset that annotates diverse labels for both tasks. Additionally, the task of annotating datasets with new labels presents significant challenges. To address this issue, a new training approach has been introduced that involves utilizing two separate datasets, one for gaze and the other for head pose. This approach allows the model to benefit from the increased amount of data from the two diverse datasets, leading to an enhanced understanding of unseen data and more robust representations for both tasks. Consequently, this results in an improvement in knowledge sharing and overall performance.

Furthermore, MTGH-Net introduces rotation matrix representation for both gaze and head pose, which is a significant advancement over traditional representations. This approach eliminates the discontinuities and ambiguities associated with previous methods and provides a continuous, unambiguous, and geometrically accurate approach for gaze estimation. Additionally, utilizing the same representation for both gaze and head pose ensures that it MTGH-Net remains unbiased and not dominated by one task over the other. Through extensive experiments on various challenging datasets, MTGH-Net has improved reliability of gaze estimation across domains and outperforms existing state-of-the-art methods in terms of generalization performance.

4.2 Key Contributions

This chapter introduces a novel approach to gaze estimation that significantly advances the state-of-the-art by addressing the critical challenge of maintaining accurate gaze estimation across different domains. The key contributions of this chapter are presented in detail:

- MTGH-Net: An innovative framework that leverages the intrinsic relationship between gaze and head pose estimation tasks through a multi-task learning

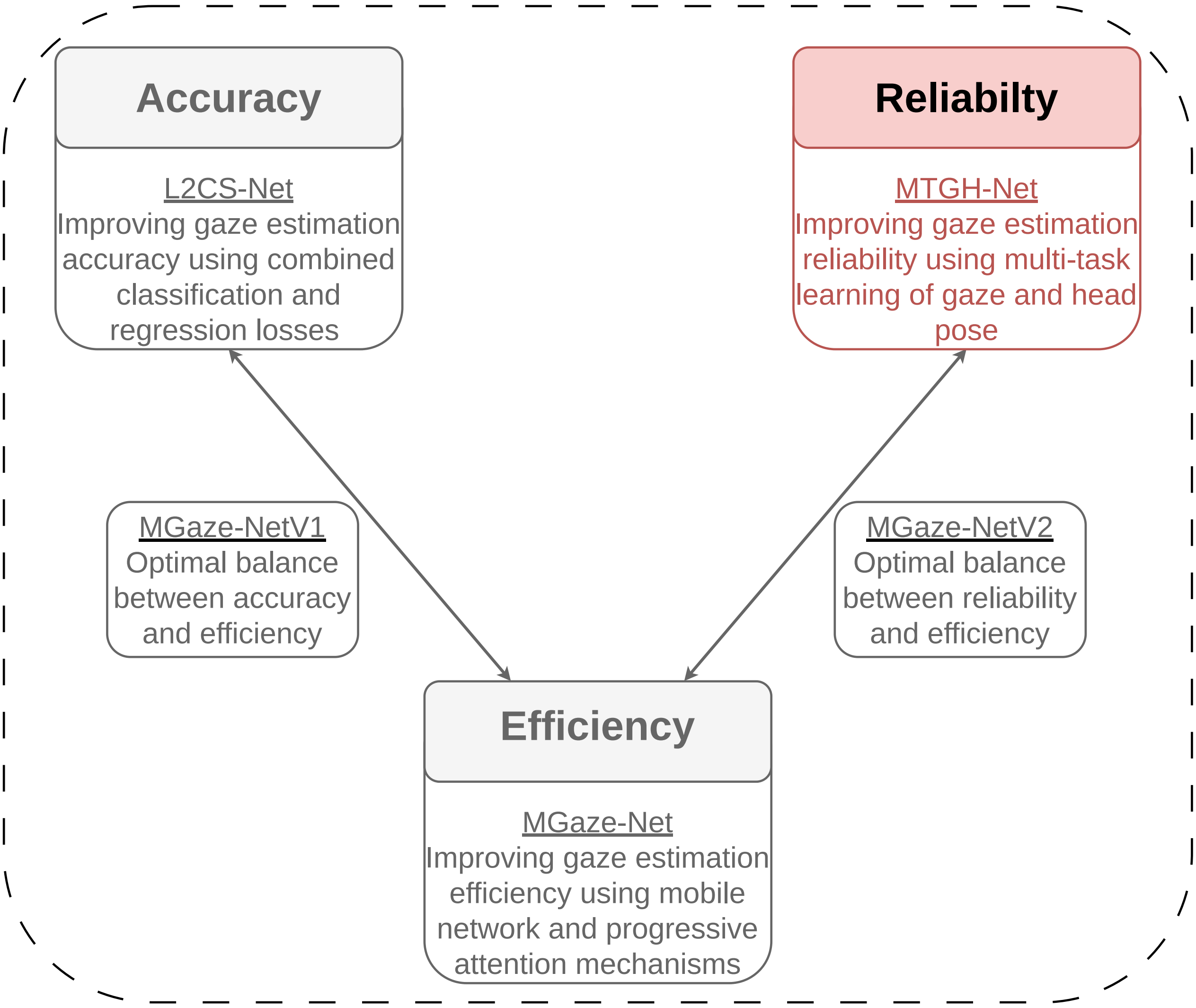

Accuracy
L2CS-Net
Improving gaze estimation accuracy using combined classification and regression losses
Reliabilty
MTGH-Net
Improving gaze estimation reliability using multi-task learning of gaze and head pose
MGaze-NetV1
Optimal balance between accuracy and efficiency
MGaze-NetV2
Optimal balance between reliability and efficiency
Efficiency
MGaze-Net
Improving gaze estimation efficiency using mobile network and progressive attention mechanisms

Figure 4.1: The goal of this chapter is to design a reliable gaze estimation model that can maintain high performance across various domains.

paradigm. This architecture harnesses the strong correlation between gaze direction and head orientation, leveraging shared and task-specific features to enhance learning efficiency and model performance. By integrating these tasks within a single network, MTGH-Net significantly reduces computational overhead and achieves synergistic improvements in estimation accuracy for gaze and head pose.

- Cross-Dataset Generalization: MTGH-Net addresses the gaze generalization challenge by utilizing a new training approach that involves utilizing two separate datasets, one for gaze and the other for head pose. This approach allows MTGH-Net to benefit from the increased amount of data from the two diverse datasets, leading to a better understanding of unseen data and more robust representations for both tasks. This strategy enriches the model's exposure to different conditions, significantly enhancing its ability to generalize across datasets with varying characteristics. This leads to an improvement in knowledge sharing and overall performance.

- Advanced Gaze and Head Pose Representation: MTGH-Net introduces a simplified and efficient 6D-parameter rotation matrix representation for both gaze and head pose estimation tasks. This representation, coupled with a geodesic-based loss function, enables accurate and direct regression of these tasks. The approach effectively overcomes the discontinuity problem inherent in traditional gaze representation methods and ensures that the model's learning process is not biased toward either task. This contribution is fundamental to the framework's ability to provide accurate estimations of gaze and head pose directions.

- Comprehensive Evaluation and Benchmarking: The detailed evaluation and benchmarking of the MTGH-Net against current state-of-the-art methods highlight its effectiveness and efficiency. The framework not only achieves SOTA performance for gaze and head pose estimation on popular benchmarks, but also demonstrates up to a 21% improvement in gaze generalization performance over single-task gaze networks. This comprehensive evaluation underscores the practical applicability and robustness of the MTGH-Net in real-world scenarios.

- Extensive Ablation Study: The inclusion of an extensive ablation study within the evaluation of MTGH-Net provides insightful analyses into the impact of its various components on overall performance. This study offers valuable insights into the contributions of the multi-task learning approach, the continuous 6D representation, and the geodesic-based loss function, among others.

The remainder of this chapter is organized as follows: Section 4.3 presents the relevant prior works. Section 4.4 presents the details of the proposed MTGH-Net. In Section 4.5, experimental, quantitative and qualitative analyses of the MTGH-Net are presented. Finally, Section 4.6 contains the conclusion.

4.3 Related Work

4.3.1 Gaze Estimation

Research in gaze estimation has seen a progression from traditional geometric models to appearance-based methods enabled by deep learning. Early approaches [122, 173, 208] predominantly focused on the use of geometric models that required explicit modeling of the eye's anatomical features, which was not only complex but also limited in accuracy and generalizability. The advent of deep learning-based gaze estimation approaches, which rely on raw images of the eye or face, advances gaze estimation field. This advancement has led to the creation of diverse datasets [54, 95, 107, 253] incorporating various environmental factors. These datasets facilitate the assessment of a model's gaze generalization capabilities, essential for practical applications. Typical gaze estimation models [27, 30, 95, 242, 251] employ CNNs to enhance estimation accuracy from single images. Cheng et al. . [30] propose a coarse-to-fine approach that refines initial gaze direction predictions from face images with eye image data. Fischer et al. [54] incorporate head pose vectors to augment gaze estimation accuracy. Yet, these models often encounter difficulties in generalizing across different domains [166] with varied appearances, lighting conditions, and head poses.

To address these challenges, domain generalization techniques have been developed, focusing on source domain data only to improve performance across diverse domains. Cheng et al. [29] innovate a gaze feature purification method using a self-adversarial framework. This way enhancing the relevance and purity of gaze features by filtering out irrelevant factors like illumination and identity. Similarly, Oh et al. [135] introduce a self-attention module that aims to refine gaze information and ensure higher resolution and more focused activation maps. Beyond domain generalization, domain adaptation strategies leverage both unlabeled and labeled data from target domains to enhance model generalization. Kothari et al. [105] improve cross-dataset accuracy by applying weakly-supervised approach to LAEO dataset (images of people looking at each other). Wang et al. [166] propose Bayesian adversarial learning for appearance feature alignment between source and target domains. Kellnhöfer et al. [95] utilize a combination of labeled and unlabeled images for model fine-tuning. Moreover, Liu et al. [116] demonstrate the use of ensemble models for collaborative learning and introduce a rotation-consistency property to enhance generalization in gaze estimation.

4.3.2 Head Pose Estimation

In recent years, the rise of neural networks has spurred advances in facial analysis and vision-based head orientation prediction. Current methods are broadly categorized into landmark-based and landmark-free approaches. Landmark-based Approaches[16, 65, 94, 108, 112, 137, 207, 268] initially detect facial landmarks and then recover the 3D head pose by aligning these landmarks with a standardized 3D head model. Under optimal conditions, this approach can yield highly accurate head orientation estimations. However, its success heavily relies on accurate landmark predictions and requires the target head to closely match the shape of the head model, which limits its applicability [180, 234, 237]. Furthermore, when facial landmarks are not visible due to significant occlusions or extreme head rotations, the effectiveness of these methods diminishes [17, 216].

landmark-free methods estimate head pose directly from images in an end-to-end manner as an appearance-based task. For instance, HopeNet [157] introduced a novel approach by binning target angle ranges and combining cross-entropy with mean squared error loss functions for Euler angle prediction. Subsequent models like QuatNet [75], HPE [80], and WHENet [263] refined this approach by varying the network architecture, using separate branches for classification and regression, or adopting different backbones like EfficientNet [174]. Recent innovations include TriNet [21] and MFDNet [115], which estimate rotation using unit vectors or rotation matrices, enhancing stability and accuracy by incorporating additional losses such as orthogonality. Probabilistic models [117] have also emerged, modeling rotation uncertainty to improve predictions (citations). Feature-focused models like FDN [235] aim to optimize head pose estimation by decoupling and specifically learning discriminative features for different head orientations (citations). DDD-Pose [4] and RankPose [41] introduce advanced augmentation schemes and ranking losses to improve training outcomes. Despite advancements, accurate estimation of head pose in unconstrained settings remained challenging due to factors such as occlusions, diverse lighting conditions, and the need for large annotated datasets that capture a wide range of head movements.

4.3.3 Multi-Task Learning

Multi-task learning (MTL) has gained recognition for its ability to enhance learning efficiency and prediction accuracy by leveraging shared representations across related tasks. This approach has found widespread application in various domains of computer vision, including action recognition [104], detection [182], and segmentation [206], as well as facial analysis [43]. Notably, MTL techniques have been successfully applied to jointly learn tasks such as facial landmark detection [108, 185], face detection [109, 224], and additional facial attributes [153, 154], underscoring the benefits of improved generalization through shared feature learning.

Despite the broad adoption of MTL techniques across these fields, their integration into gaze estimation has been relatively limited. Ghosh et al. [60] introduced a novel multi-task gaze estimation framework that incorporates head pose estimation as a supporting task. This approach utilizes an existing model to generate the necessary head pose labels for the multi-task network, effectively reducing variability caused by changes in head pose. However, this method introduces challenges, including the need for additional labeling of head pose information within gaze datasets, which could lead to inaccuracies or require significant time and effort. Lian et al. [113] propose a novel multitask method for gaze point estimation using multiview cameras. This method delves into the intricate relationship between gaze point and gaze direction estimation, utilizing a partially shared CNN architecture to estimate both metrics simultaneously. Xue et al. [127] leverage the least absolute deviation (LAD) and utilize visual features like intensity, color, and texture in a multitask multiview sparse learning approach. The approach is integrated into a particle filter framework, where sparse representations for each view of a particle are treated as individual tasks, exploiting the relationships between different views and particles through a robust multitask formulation. Yu et al.[230] have exploited the intrinsic link between gaze direction and eye landmark positions, employing an MTL framework to predict both sets of

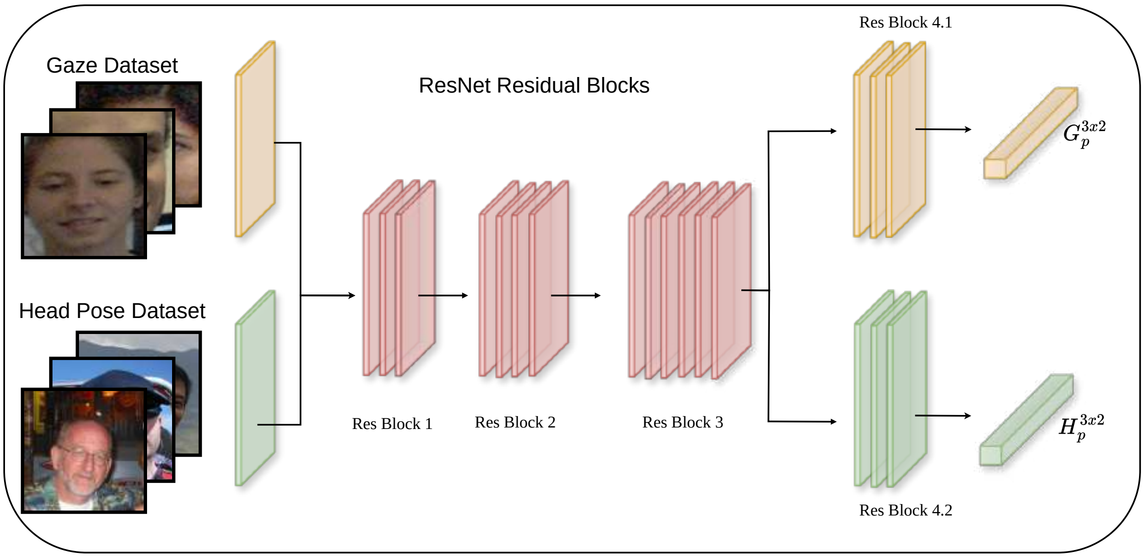


Figure 4.2: Overview of the proposed MTGH-Net, which consists of a modified ResNet-50 network as the bottom layer and two fully connected layers as the top layers.

data concurrently. This strategy not only highlights the potential for MTL techniques to streamline gaze estimation but also enhances the overall process by simultaneously addressing related tasks. The integration of these related tasks within a single learning framework exemplifies the synergy possible with MTL, promising more robust and accurate models in the field of gaze estimation.

4.4 Framework

Accurate gaze estimation in different domains depends on several factors, including head pose information and the availability of large-scale datasets with various unconstrained settings. Head pose plays a crucial role in estimating gaze direction, as there is a strong interplay between eye and head [145, 249, 258]. Additionally, creating a large-scale dataset with diverse unconstrained settings is important for robust and reliable gaze estimation. Traditional methods, which often rely on limited gaze-annotated data and overlook the complexity of integrating head pose, face challenges in achieving high accuracy and generalizability. The proposed MTGH-Net utilizes a novel multi-task learning architecture designed to efficiently and accurately estimate both gaze and head pose. By distilling knowledge from extensive datasets that encompass both gaze and head pose information, MTGH-Net aims to improve accuracy and improve generalizability across different domains. This section elaborates on the architecture, components, and functionalities of the MTGH-Net. A complete pipeline of the proposed MTGH-Net is depicted in Figure 4.3.

4.4.1 Network Architecture

The architecture of the MTGH-Net network, detailed in Fig. 4.3, is based on a modified ResNet-50 backbone, followed by two fully connected layers for tasks. The input to the network includes facial images from gaze and head pose datasets. At the core of the MTGH-Net is a modified version of the ResNet-50 architecture, which serves as the backbone for feature extraction. The selection of ResNet-50 is based on its proven effectiveness in various computer vision tasks, as it offers deep convolutional layers that are adept at capturing complex visual patterns from images. The standard ResNet-50 architecture, known for its convolutional layer, four residual blocks, and a global average pooling layer, is modified to better serve the requirements of MTGH-Net approach. Specifically, MTGH-Net is divided into bottom layers and top layers. The bottom layers of the network consist of the initial convolutional layer and the first three residual blocks of the standard ResNet-50 architecture. These layers are responsible for extracting common features from the input images that are relevant to both tasks. This shared layers encodes input data into a common feature representation, facilitating the modeling of intricate relationships between both tasks.

The top layers of the network consist of the remaining residual block of the ResNet-50 architecture, followed by the global average pooling and fully connected layers, which culminate in the final output predictions of gaze and head pose, respectively. After the joint feature extraction phase in the bottom layers, the architecture is split into two separate pipelines, each designed for each task gaze and head pose. This separation allows the network to customize the extraction of features relevant to each task, thus refining initial shared representations into more task-specific features. This dual branch design is pivotal since it maintains the integrity of task-specific features by processing them through task specific layers, ensuring that each task's unique features are preserved and accurately designed.

4.4.2 Unified Representation

In multitask learning scenarios, different representations of tasks can lead to imbalanced losses. These inconsistencies can cause one task to dominate the learning process and reduce the influence of the second task on training. In this approach, gaze and head pose use two distinct representations, e.g., Euler angels for head pose and spherical angles for gaze which may scale the losses differently. Further, these representation present a barrier for each task since they suffer from ambiguity and discontinuity which can lead to decreased estimation performance. By adopting a consistent representation for both tasks, MTGH-Net ensures that losses are scaled equally, facilitating a more balanced and effective learning process.

A key contribution of MTGH-Net is the adoption of a rotation matrix representation for gaze and head pose to guarantee a unified representation for both tasks and to address the challenges associated with traditional representation. This representation

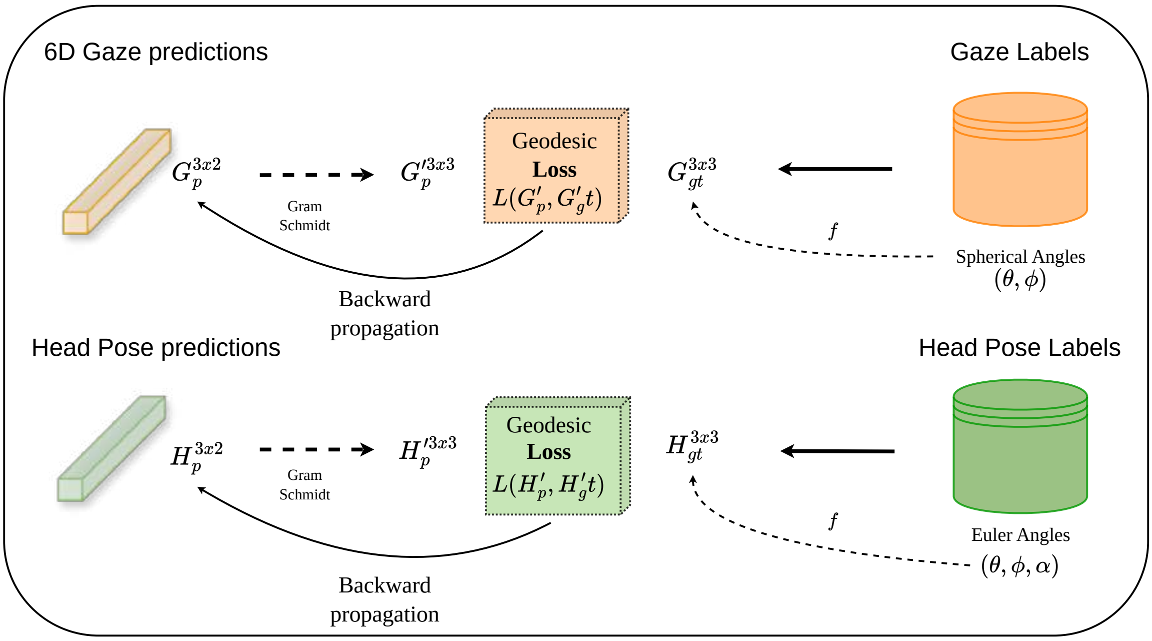


Figure 4.3: MTGH-Net utilizes a gaze representation based on the rotation matrix using Gram-Schmidt and a geodesic-based loss function for both gaze and head pose.

allows for direct regression of gaze and head pose orientations from the shared feature space, overcoming discontinuities and biases introduced by other representation. The 6D representation is achieved by predicting two orthogonal vectors that define the gaze direction and head orientation in 3D space, which are then processed through the Gram-Schmidt orthogonalization to ensure the output conforms to the rotation matrix format. This approach facilitates a more accurate and robust estimation process.

4.4.2.1 Gaze Rotation Representation

This section delves into the representation of gaze direction using rotation matrices, offering a continuous and unambiguous parameterization. In the domain of computer vision, accurately representing gaze direction is pivotal for various applications. The early methods [55] in gaze estimation predominantly utilized 2D pixel coordinates, denoted by $\mathbf{g} = \{x, y\}$, to define the point of gaze fixation on a screen, with x and y representing the horizontal and vertical coordinates, respectively. This approach, while commonly used in human-computer interaction scenarios, is insufficient for tasks that require a three-dimensional understanding of gaze direction to accurately capture human attention and intention [3, 192, 193]. To bridge this gap, advancements in gaze estimation have introduced the concept of a unit gaze vector [55, 265], denoted as $\mathbf{g} = \{\mathbf{o}, \mathbf{v}\}$ where $\mathbf{o} \in \mathbb{R}^3$ is the origin of the gaze vector and $\mathbf{v} \in \mathbb{R}^3$ represents a normalized vector pointing in the direction of gaze. This representation shifts the paradigm to a 3D context, offering a more detailed and accurate depiction of gaze orientation, essential for real-world interaction applications.

Recently, the common practice has been to use spherical coordinates(θ, ϕ) [14, 27, 30, 54, 95, 251, 253] to define the 3D gaze direction in the eye coordinate system. This approach maps the 3D unit gaze vector $\mathbf{g} = [g_x, g_y, g_z]^\intercal$ to azimuth $\theta = \arcsin(g_y)$ and elevation $\phi = \arctan(g_x, g_z)$ angles, providing a compact and intuitive means of describing gaze direction. However, it is important to note that while spherical coordinates simplify the representation of gaze direction and reduce output space dimensionality, they also introduce challenges related to discontinuity and representation ambiguity. Overcoming these challenges requires continuous and unambiguous representations to enhance the learning process and the accuracy of gaze estimation models. The rotation matrix $R^{3\times 3}$ offers a comprehensive means of representing rotations, providing a continuous and uniquely parameterized representation for each possible rotation. This approach is particularly advantageous for gaze estimation, where the need for accurate and unambiguous descriptions of eye orientations is paramount. The initial orientation of the eyes, assumed to be frontal, can be denoted in Euler angles as $(0, 0)$, reflecting zero azimuth and elevation. The transition from this baseline orientation to a specific gaze direction can then be captured using the angles (θ, ϕ) in the euler angels notation. These angles can be transformed into a rotation matrix that can represents the direction of gaze as follows:

For the vertical rotation about x-axis by angle (θ), the corresponding rotation matrix $R_x(\theta)$ is:

$$R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{4.1}$$

For the horizontal rotation about y-axis by angle ϕ, the rotation matrix $R_y(\phi)$ is defined as:

$$R_y(\phi) = \begin{bmatrix} \cos(\phi) & 0 & \sin(\phi) \\ 0 & 1 & 0 \\ -\sin(\phi) & 0 & \cos(\phi) \end{bmatrix} \tag{4.2}$$

The gaze can be converted using (θ, ϕ) to rotation matrix by combining these rotations that yields the full rotation matrix representing the eyes' orientation:

$$R_{gaze}(\theta, \phi) = R_y(\phi) \cdot R_x(\theta). \tag{4.3}$$

4.4.2.2 Head pose Rotation Representation

Head pose estimation is crucial in various fields such as augmented reality, robotics, and driver assistance systems. A common method to represent head orientation is by using Euler angles, which describe the orientation of the head in terms of three rotations about the axes of a coordinate system. This section provides a detailed derivation of converting Euler angles roll (ϕ), pitch (θ), and yaw (ψ) into a rotation matrix. Euler angles consist of three angles includes yaw (ψ) rotation about the x axis, pitch (θ) rotation about the y axis, and roll (ϕ) rotation about the z axis.

These rotations are applied sequentially to derive the overall orientation of the head in a three-dimensional space. Each rotation can be represented as a matrix:

- Yaw (ψ):

$$R_x(\psi) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\psi) & -\sin(\psi) \\ 0 & \sin(\psi) & \cos(\psi) \end{bmatrix}$$

- Pitch (θ):

$$R_y(\theta) = \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{bmatrix}$$

- Roll (ϕ):

$$R_z(\phi) = \begin{bmatrix} \cos(\phi) & -\sin(\phi) & 0 \\ \sin(\phi) & \cos(\phi) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The overall rotation matrix R that represents the head pose is obtained by multiplying the individual rotation matrices. The multiplication order depends on the specific application and defines how these rotations interact:

$$R_{head}(\psi, \theta, \phi) = R_z(\phi) \cdot R_y(\theta) \cdot R_x(\psi)$$

This multiplication represents the composition of rotations starting with yaw, followed by pitch, and finally roll.

This final rotation matrix R encapsulates the entire head pose in a continuous, non-ambiguous form, overcoming the limitations associated with using Euler angles directly. This method of deriving the rotation matrix from Euler angles provides a robust and clear framework for representing 3D orientations, crucial for accurate head pose estimation in various applications.

4.4.2.3 6D Parameter Representation

To effectively predict gaze and head pose directions, MTGH-Net uses a rotation matrix representation for each task, aiming to predict a total of 9 values. However, directly predicting these values introduces significant challenges, notably ensuring that the column vectors of the matrix remain orthogonal and the determinant equals 1. Inspired by recent advances in rotation matrix representation, a 6D representation strategy is adapted for the rotation matrix involved in gaze and head pose estimations. This approach involves predicting only two 1×3, vectors $\boldsymbol{a_1}$ and $\boldsymbol{a_2}$ effectively reducing the output dimensionality to 6D. The formulation is as follows:

$$\boldsymbol{g_{GS}} = \begin{bmatrix} \boldsymbol{a_1} & \boldsymbol{a_2} \end{bmatrix}. \tag{4.4}$$

This simplifies the full rotation matrix representation by excluding the last column vector $\boldsymbol{a_3}$. To ensure the orthogonality constraint and maintain a valid rotation matrix, the Gram-Schmidt orthogonalization process is employed. This method allows for the construction of the missing third vector, $\boldsymbol{b_3}$ and ensures that all column vectors are orthogonal, and the determinant of the rotation matrix is 1. The vectors $\boldsymbol{b_1}$, $\boldsymbol{b_2}$ and $\boldsymbol{b_3}$ are defined as:

$$\begin{aligned} \boldsymbol{b_1} &= N(\boldsymbol{a_1}), \\ \boldsymbol{b_2} &= N(\boldsymbol{a_2} - (\boldsymbol{b_1} \cdot \boldsymbol{a_2})\boldsymbol{b_1}), \\ \boldsymbol{b_3} &= \boldsymbol{b_1} \times \boldsymbol{b_2}. \end{aligned} \tag{4.5}$$

where $N(\cdot)$ denotes a normalization function and $\boldsymbol{b_3}$ is computed via the cross product of $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$ to satisfy orthogonality constrains. By employing this 6D representation and the subsequent transformation process, MTGH-Net simplifies the prediction of gaze and head pose directions. This approach not only circumvents the direct regression of the full rotation matrix but also seamlessly integrates the orthogonality and determinant constraints, enhancing the model's learning efficacy and the accuracy of its predictions.

4.4.3 Training Approach

MTGH-Net introduces a novel training strategy designed to leverage the synergy between gaze and head pose estimation tasks while addressing the inherent challenges of cross-dataset generalization. This section provides an in-depth analysis of the MTGH-Net's training strategy, emphasizing its innovative approach to dataset utilization, loss function integration, and optimization techniques.

4.4.3.1 Dual Dataset Training

In the realm of multi-task learning (MTL), the conventional strategy relies on utilizing a unified dataset annotated with ground truth for each task. However, a significant challenge arises in designing MTGH-Net due to the absence of a large scale dataset that encompasses annotations for both tasks, especially with diverse values. Furthermore, the process of annotating existing datasets with new labels presents considerable accuracy challenges. To address these challenges, MTGH-Net utilizes two distinct datasets for training, one with gaze labels and the other with head pose annotation. Training MTGH-Net on separate but complementary datasets allows access to a broader

Algorithm 1: The Proposed Face Tracking Algorithm

```
Inputs : $\boldsymbol{g}(\theta, \phi)$, ground truth in spherical coordinates for gaze annotated
         dataset $\mathcal{D}_g = (\{\boldsymbol{x}_i, \boldsymbol{g}_i\}_{i=1}^{n_g})$ and $\boldsymbol{h}(\theta, \phi, \alpha)$, ground truth in Euler angles
         for head pose annotated dataset $\mathcal{D}_h = (\{\boldsymbol{y}_i, \boldsymbol{h}_i\}_{i=1}^{n_h})$.
Output: $\mathcal{F}_f$, $\mathcal{F}_g$, $\mathcal{F}_h$
$G_{gt}^{3\times3} \leftarrow eul2rotm(g(\theta, \phi))$;
$H_{gt}^{3\times3} \leftarrow eul2rotm(h(\theta, \phi, \alpha))$;
foreach $e \in epochs$ do
    $f \leftarrow \mathcal{F}_f(x; \theta_f)$ ;
    $G_{6d}^{3\times2} \leftarrow \mathcal{F}_g(f; \theta_g)$ ;
    $G_p^{3\times3} \leftarrow GramSchmidt(G_{6d}^{3\times2})$;
    $\mathcal{L}_g(\theta_f, \theta_g) = \arccos\left(\frac{tr(G_p G_{gt}^T)-1}{2}\right)$;
    $\{\theta_f, \theta_g\} \leftarrow \bigtriangledown_{\{\theta_f, \theta_g\}} \mathcal{L}_g$;
    $f \leftarrow \mathcal{F}_f(y; \theta_f)$ ;
    $H_{6d}^{3\times2} \leftarrow \mathcal{F}_h(f; \theta_h)$ ;
    $H_p^{3\times3} \leftarrow GramSchmidt(H_{6d}^{3\times2})$;
    $\mathcal{L}_h(\theta_f, \theta_h) = \arccos\left(\frac{tr(H_p H_{gt}^T)-1}{2}\right)$;
    $\{\theta_f, \theta_h\} \leftarrow \bigtriangledown_{\{\theta_f, \theta_h\}} \mathcal{L}_h$;
end
```

spectrum of data, which includes a wide variety of head poses and gaze directions across diverse environments and demographic groups. This exposure to diverse data significantly enhances the framework's ability to generalize across unseen datasets, marking a major advancement over previous methods that often suffered from overfitting to specific dataset features. Furthermore, it ensures that the learned model does not overfit the single dataset and is robust enough to perform accurately across unseen datasets. MTGH-Net employs a unique training strategy that each batch in the training process is a tuple containing samples for head pose and gaze, which are processed consecutively in an alternating fashion. These samples are initially passed through the shared bottom layers of the network to leverage the correlation between the two tasks. Subsequently, they are routed exclusively through their respective top layer branches designed for either head pose or gaze. This architecture allows the bottom layers to learn general features relevant to both tasks, while the top branches specialize in task-specific processing. During testing, input images are passed through both branches of the network to generate predictions for both head pose and gaze.

The training paradigm of MTGH-Net is illustrated in algorithm 1. Two distinct datasets are utilized: a gaze-annotated dataset, D_g, and a head pose-annotated dataset, D_h. The gaze-annotated dataset, denoted by $\mathcal{D}_g = \left\{(\boldsymbol{x}_i, \boldsymbol{g}_i)|_{i=1}^{n_g}\right\}$, includes n_g image label pairs. Similarly, the head pose-annotated dataset is referenced by $\mathcal{D}_h =$

$\{(\boldsymbol{y}_i, \boldsymbol{h}_i)|_{i=1}^{n_h}\}$ and includes n_h image label pairs. The MTGH-Net, denoted as $\mathcal{F}_{(\theta_f,\theta_h,\theta_g)}$, integrates several key components. It incorporates θ_f, which are shared feature embedding parameters for processing input from both datasets. Additionally, θ_g represents gaze-specific top layer parameters for gaze estimation. Lastly, θ_h includes head pose-specific top layer parameters for head pose estimation.

For gaze estimation, input image $\boldsymbol{x} \in \mathbb{R}^{224\times 224}$ from the gaze-annotated dataset are mapped to the shared feature embedding space $\boldsymbol{f} \in \mathbb{R}^{2048}$ by $\mathcal{F}_f : \boldsymbol{x} \to \boldsymbol{f}$. The features $\boldsymbol{f}$ are then mapped to a 6D gaze output representation, $\boldsymbol{G_{6d}} \epsilon \mathbb{R}^{3\times 2}$ by $\mathcal{F}_g : \boldsymbol{f} \to \boldsymbol{G_{6d}}$. The 6D gaze representation is further processed into a 3×3 rotation matrix, $\boldsymbol{G_p} \in \mathbb{R}^{3\times 3}$ adhering to the orthogonality constraint via the Gram-Schmidt process.

For head pose estimation, input image $\boldsymbol{y} \in \mathbb{R}^{224\times 224}$ from the annotated head pose dataset are mapped to the shared feature embedding space $\boldsymbol{f} \in \mathbb{R}^{2048}$ by $\mathcal{F}_f : \boldsymbol{y} \to \boldsymbol{f}$, and then to a head pose output space $\boldsymbol{H_{6d}} \epsilon \mathbb{R}^{3\times 2}$ by $\mathcal{F}_h : \boldsymbol{f} \to \boldsymbol{H_{6d}}$. The 6D head representation is further processed into a 3×3 rotation matrix, $\boldsymbol{H_p} \in \mathbb{R}^{3\times 3}$ adhering to the orthogonality constraint via the Gram-Schmidt process.

This dual-dataset training strategy ensures that while the shared bottom layers learn to extract features relevant to both gaze and head pose, the task-specific top layers refine these features for accurate task-specific predictions. This architecture not only facilitates efficient training across tasks but also maximizes the performance and generalization capabilities of the MTGH-Net model.

4.4.3.2 Geodesic Loss Function

For gaze estimation, the geodesic error between the predicted rotation matrix G_p and G_{gt} is computed as follows:

$$\mathcal{L}_g(\theta_f, \theta_g) = cos^{-1}\left(\frac{tr(G_p G_{gt}^T) - 1}{2}\right). \tag{4.6}$$

Similarly, for estimation of the head pose, the geodesic error between the predicted rotation matrix H_p and H_{gt} is defined by:

$$\mathcal{L}_h(\theta_f, \theta_h) = cos^{-1}\left(\frac{tr(H_p H_{gt}^T) - 1}{2}\right). \tag{4.7}$$

4.5 Experiments and Results

MTGH-Net undergoes extensive experimentation to validate its effectiveness and superiority over existing state-of-the-art methods. This section elaborates on the experimental setup, the datasets used for evaluation, and the results obtained, highlighting the achievements of MTGH-Net in both gaze and head pose estimation tasks.

4.5.1 Experimental Protocols

This section includes the experimental protocols required for assessing the proposed framework including evaluation datasets, performance metrics, data preprocessing, implementation details, and experiments setup.

4.5.1.1 Evaluation Datasets

MTGH-Net is tested against multiple public datasets known for their diversity in subject demographics, environmental settings, and capture conditions. These include MPIIFaceGaze, GazeCapture, Gaze360, and RT-GENE for gaze estimation, and 300W-LP, AFLW2000, and BIWI for head pose estimation. Each dataset presents unique challenges, from variations in lighting and background to differences in resolution and head pose distribution, making them ideal for evaluating the MTGH-Net generalizability.

Gaze annotated datasets: Four popular gaze benchmarks are employed in MTGH-Net consisting of MPIIFaceGaze, GazeCapture, Gaze360, and RT-GENE. The gaze distribution plots of all these datasets and their example face images are shown in Fig. 4.4.

- **MPIIFaceGaze** [248]: It offers 213,659 images from 15 subjects, captured over several months of their daily activities. The wide variety of backgrounds and lighting conditions in this dataset makes it particularly suited for unconstrained gaze estimation tasks.

- **GazeCapture** [107]: is the largest in-the-wild gaze dataset currently available. It includes a vast array of images: 1,379,083 for training, 191,842 for testing, and 63,518 for validation. Its extensive size and real-world applicability make it an invaluable asset for evaluating robust gaze estimation models.

- **RT-GENE** [54]: offers a detailed collection of 122,531 samples from 15 subjects that captured using wearable eye-tracking glasses. This indoor dataset is characterized by its high variance in both gaze and head pose angles, offering a challenging and valuable resource for gaze estimation studies.

- **Gaze360** [95]: is notable for its diversity, featuring images from 238 subjects across a wide range of ages, genders, and ethnic backgrounds. It includes both indoor and outdoor settings which offer a comprehensive view of natural gaze behavior. From this dataset, 84,900 images featuring frontal faces are specifically utilized to ensure consistency in the analysis. For consistency, 84,900 images featuring frontal faces are specifically utilized.

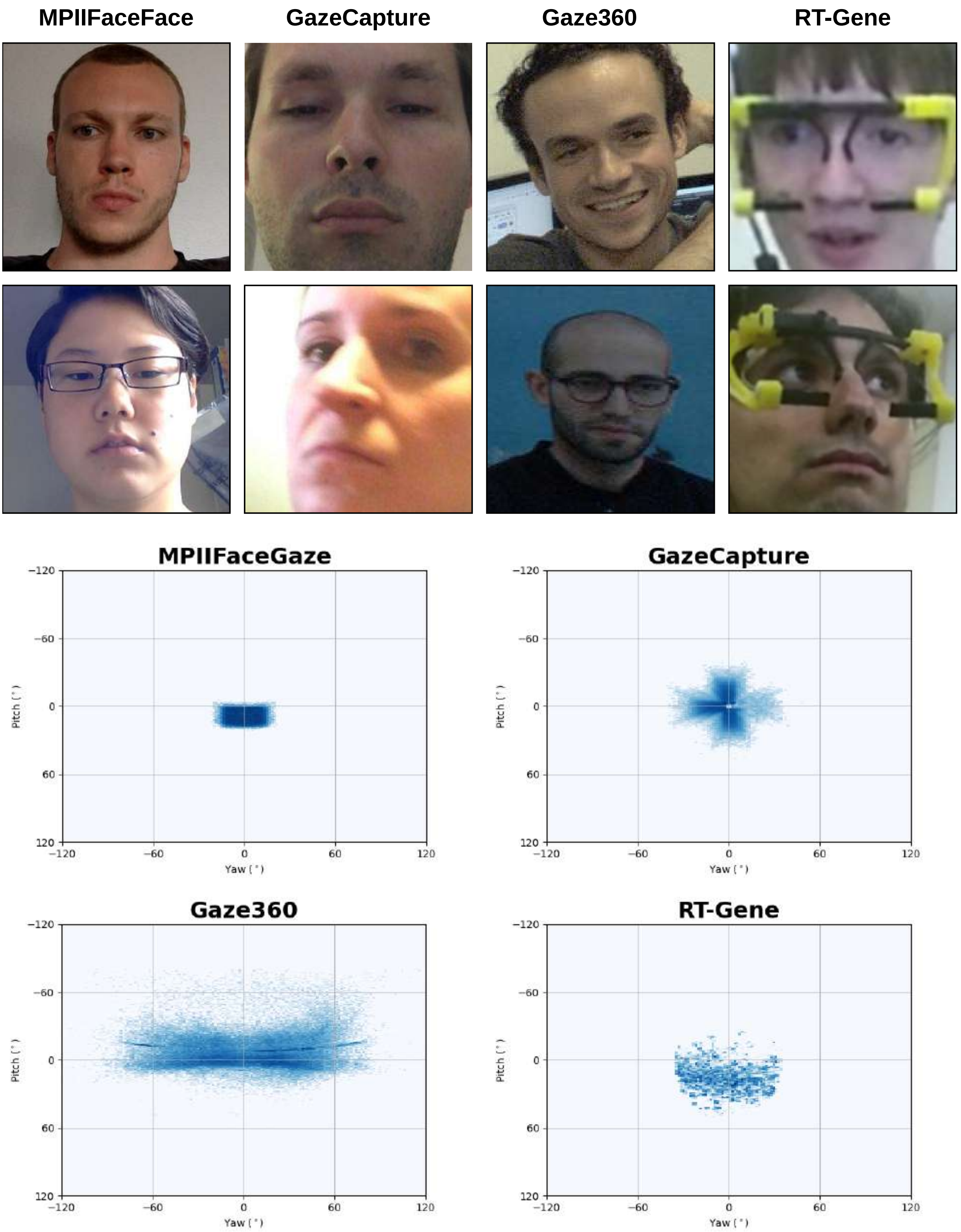


Figure 4.4: Illustration of the gaze direction distributions of the four gaze datasets. The upper row shows the image samples from the four gaze datasets. The bottom row is the gaze direction distribution statistics.

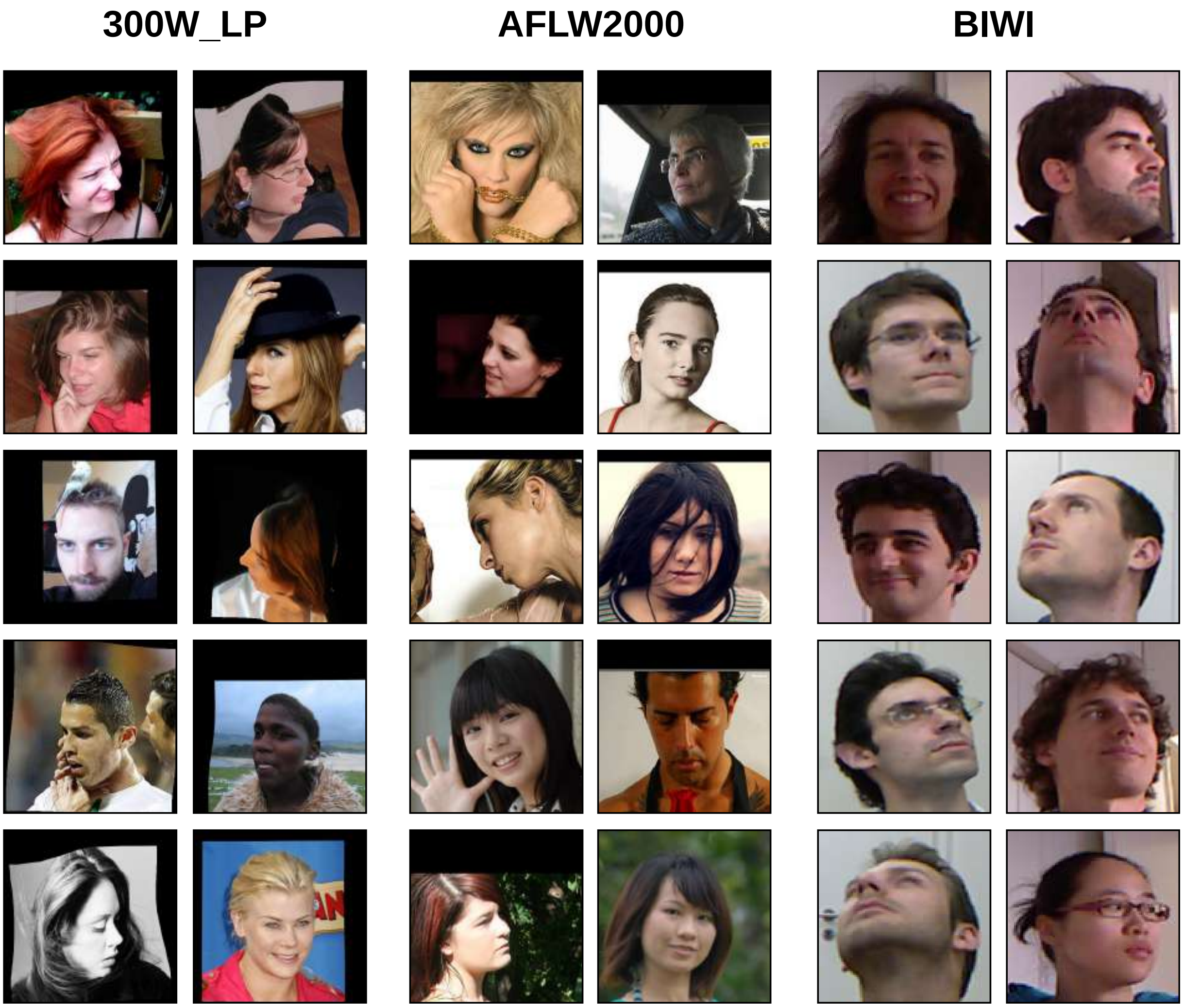


Figure 4.5: Face image samples from the three head pose datasets with diversity of environmental conditions, illuminations and head poses.

For Gaze360, RT-Gene and MPIIFaceGaze, the preprocessing protocols outlined in previous studies [34, 245] are utilized which focus on creating normalized face crops to facilitate consistent gaze estimation. For the GazeCapture dataset, the preprocessing and normalization techniques described in [231, 245] are applied to generate standardized head crops, ensuring uniformity across all datasets.

Head pose annotated datasets: MTGH-Net incorporates three widely recognized head pose benchmark datasets. These datasets are chosen for their comprehensive coverage of head poses and their contribution to the diversity of training and validation samples.

- **300W-LP** [267]: combines multiple standard facial landmark annotated datasets, including AFW [266], LFPW [13], HELEN [260] and IBUG [160]. It utilizes facial profiling techniques to generate a wide range of synthesized head poses.

It encompasses 61,225 images, which are effectively doubled to 122,450 images through image flipping, offering a broad spectrum of head poses for training.

- **AFLW2000** [269]: comprises 2,000 images selected for their variety and the challenging nature of their head poses. Despite its relatively small size, it provides a potent validation dataset due to the complexity and diversity of the head poses it features.
- **BIWI** [53]: contains 15,678 images from 20 subjects, captured in an indoor setting. Its focus on a controlled environment allows for detailed study of head pose variations across different subjects, further enriching the data diversity for MTGH-Net.

By incorporating these datasets, MTGH-Net is equipped with a rich variety of head pose annotations, enabling it to learn and predict across a wide range of real-world conditions.

4.5.1.2 Performance Metrics

To assess the accuracy and reliability of MTGH-Net, distinct evaluation metrics are employed for gaze and head pose estimation tasks, reflecting the specific challenges and requirements of each task.

For gaze estimation, the Mean Angular Error (MAE) serves as the primary metric. It quantifies the discrepancy between the ground-truth gaze direction, g and the predicted gaze direction ĝ, in terms of their angular difference. The MAE is calculated as follows:

$$\textsc{Angular Error} = \arccos \frac{\mathbf{g} \cdot \hat{\mathbf{g}}}{||\mathbf{g}||\,||\hat{\mathbf{g}}||} \tag{4.8}$$

For the head pose estimation task, the Mean Absolute Error (MAE) is adopted as the evaluation metric. This metric calculates the average absolute difference between the ground-truth pose parameters, $\boldsymbol{x_g}$ and the predicted head pose values $\boldsymbol{x_p}$, across all N images in the dataset:

$$\textsc{Mean Absolute Error} = \frac{1}{N} \sum_{i=1}^{N} (|\boldsymbol{x_g} - \boldsymbol{x_p}|) \tag{4.9}$$

4.5.1.3 Implementation Details

The experiments are conducted using the PyTorch framework (version 1.8.1) on a state-of-the-art computing setup equipped with an Intel(R) Core (TM) i7-7800X CPU and an NVIDIA RTX 3080 with 12GB of memory. MTGH-Net is based on a ResNet50 convolutional backbone and utilizes pre-trained weights on the extensive ImageNet-1K dataset. [159]. Due to the disparity in the number of images available in the gaze and head pose datasets, dataset-specific batch sizes are employed: 50 for the head pose dataset, 72 for the $\mathcal{D}_{GZ}$ dataset, and 66 for the $\mathcal{D}_{RT}$ dataset. The network undergoes training for a total of 50 epochs, utilizing the Adam optimizer to facilitate efficient convergence with the learning rate of $1e^{-5}$. To maintain consistency and ensure that the network effectively learns relevant features, images from both gaze and head pose datasets undergo cropping and resizing to a uniform dimension of 224×224 pixels.

4.5.1.4 Experiments Setup

To assess the reliability of MTGH-Net, two experiments are conducted that focus on the evaluation of the data set between the gaze and the head pose. The first experiment includes training MTGH-Net simultaneously on Gaze360 ($\mathcal{D}_{GZ}$) and 300W-LP dataset ($\mathcal{D}_{WL}$) datasets. $\mathcal{D}_{GZ}$ is selected as it has the widest range of gaze values in all datasets. $\mathcal{D}_{WL}$ is chosen since it is the most widely used dataset for head pose estimation. The second experiment include training MTGH-Net simultaneously on RT-Gene ($\mathcal{D}_{RT}$) and 300W-LP dataset ($\mathcal{D}_{WL}$) datasets. $\mathcal{D}_{RT}$ is chosen for its narrower range of gaze angles to assess the performance of MTGH-Net under more constrained conditions. Consequently, four cross-dataset evaluation tasks are established include:

- $\boldsymbol{\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}}$: In this task, MTGH-Net is trained simultaneously on Gaze360 ($\mathcal{D}_{GZ}$) and 300W-LP dataset ($\mathcal{D}_{WL}$) and tested on MPIIFaceGaze ($\mathcal{D}_{MP}$).

- $\boldsymbol{\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}}$: In this task, MTGH-Net is trained simultaneously on Gaze360 ($\mathcal{D}_{GZ}$) and 300W-LP dataset ($\mathcal{D}_{WL}$) and tested on GazeCapture ($\mathcal{D}_{GC}$).

- $\boldsymbol{\mathcal{D}_{RT} \rightarrow \mathcal{D}_{MP}}$: In this task, MTGH-Net is trained simultaneously on RT-GENE ($\mathcal{D}_{RT}$) and 300W-LP dataset ($\mathcal{D}_{WL}$) and tested on MPIIFaceGaze ($\mathcal{D}_{MP}$).

- $\boldsymbol{\mathcal{D}_{RT} \rightarrow \mathcal{D}_{GC}}$: In this task, MTGH-Net is trained simultaneously on RT-GENE ($\mathcal{D}_{RT}$) and 300W-LP dataset ($\mathcal{D}_{WL}$) and tested on GazeCapture ($\mathcal{D}_{GC}$).

The tasks $\boldsymbol{\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}}$ and $\boldsymbol{\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}}$ are used to assess the generalization performance of MTGH-Net trained on the Gaze360 dataset to the MPIIFaceGaze and GazeCapture datasets, respectively. Similarly, tasks $\boldsymbol{\mathcal{D}_{RT} \rightarrow \mathcal{D}_{MP}}$ and $\boldsymbol{\mathcal{D}_{RT} \rightarrow \mathcal{D}_{GC}}$ are employed to evaluate MTGH-Net when trained on the RT-GENE dataset and tested on the MPIIFaceGaze and GazeCapture datasets. Additionally, two cross-dataset head pose evaluation tasks are formulated to further examine the generalization performance of MTGH-Net across different head pose domains. These tasks, $\boldsymbol{\mathcal{D}_{WL} \rightarrow \mathcal{D}_{AF}}$ and $\boldsymbol{\mathcal{D}_{WL} \rightarrow \mathcal{D}_{BI}}$, test the adaptability of the MTGH-Net network trained on the 300W-LP dataset to the AFLW2000 and BIWI datasets, focusing on the model's capacity to handle diverse head pose domains.

Table 4.1: Performance comparison with state of the art methods on the two gaze cross-dataset tasks include $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}$ and $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}$. MTGH-Net shows the best performance among typical gaze estimation methods and domain generalization approaches.

Category	Methods	Multi-Task	$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}$	$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}$
Typical gaze estimation Methods	Full-Face [251]	NO	13.53°	22.23°
	RTGene [54]	NO	14.52°	20.11°
	CA-Net [30]	NO	16.27°	22.11°
	Dilated-Net [27]	NO	11.72°	18.9°
	Eth-Gaze [242]	NO	10.3°	**16.98°**
	Gaze360 [95]	NO	**8.15°**	17.80°
Domain Generalization Methods	PureGaze [29]	NO	9.28°	**17.22°**
	SAtten-Net [135]	NO	**8.31°**	17.70°
	$\boldsymbol{MTGH-Net_G}$	**YES**	**6.0°**	**15.50°**

4.5.2 Quantitative Results

This section presents the quantitative results of the experiments, demonstrating the superiority of MTGH-Net in improving reliability of gaze estimation over existing state-of-the-art techniques in cross dataset evaluation.

4.5.2.1 Cross Dataset Evaluation

MTGH-Net is benchmarked against both typical gaze estimation methods and domain generalization approaches. Typical gaze estimation methods such as Full-Face [251], RT-Gene [54], CA-Net [30], Dilated-Net [27], Eth-Gaze [242], and Gaze360 [95]. These methods generally aim to enhance within-dataset accuracy through comprehensive network designs and approaches. However, they often fail in cross-dataset scenarios due to the overfitting to specific domain characteristics. In contrast, domain generalization approaches such as SAtten-Net [135] and PureGaze [29], designed to enhance cross-dataset performance without incorporating target domain samples. These approaches generally exhibit improved reliability over typical gaze estimation methods, but still struggle to completely overcome the limitations imposed by source domain bias. Table 4.1 and Table 4.2 present a comparative analysis of mean angular error, comparing MTGH-Net with existing state-of-the-art typical gaze estimation methods and domain generalization on the four cross dataset tasks. The top three best gaze estimation performance results are highlighted in bold in the tables.

From Table 4.1, typical gaze estimation methods such as Full-Face [251], CA-Net [30] and RT-Gene [54] exhibit higher angular errors, indicating less adaptability to unseen domains. In contrast, domain generalization methods like PureGaze [29] and SAtten-Net [135] show better performance than most typical methods but still do not outperform Gaze360[95]and Eth-Gaze [242]. Gaze360, which utilizes a 7-frames

Table 4.2: Performance comparison with SOTA methods on two gaze cross-dataset tasks. MTGH-Net shows the state of the art performance among typical gaze estimation methods and domain generalization approaches.

Category	**Methods**	**Multi-Task**	$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{MP}$	$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{GC}$
Typical gaze estimation Methods	Full-Face [251]	NO	14.40°	18.16°
	RTGene [54]	NO	10.95°	15.62 °
	CA-Net [30]	NO	15.62°	18.23°
	Dilated-Net [27]	NO	**8.92°**	14.21°
	Eth-Gaze [242]	NO	12.0°	**13.20°**
	Gaze360 [95]	NO	**9.12°**	**13.58°**
Domain Generalization Methods	PureGaze [29]	NO	-	-
	SAtten-Net [135]	NO	-	-
	$\boldsymbol{MTGH-Net_R}$	**YES**	**8.29°**	**11.41°**

LSTM network to predict gaze values, achieves the best performance among typical methods, particularly on the $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}$ task with an error of 8.15°. Moreover, Eth-Gaze [242] demonstrates relatively low angular error, especially in the $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}$ task with 16.98°. However, MTGH-Net reports the lowest angular error of 6.00° and 15.50° in $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}$ and $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}$ tasks respectively. MTGH-Net outperforms the last reported state-of-the-art of Gaze360[95] by approximately 27% and surpasses Eth-Gaze by approximately 9%.

From Table 4.2, Dilated-Net [27] achieves the lowest error among typical gaze estimation methods with 8.92° in the $\mathcal{D}RT \rightarrow \mathcal{D}MP$ task, while Eth-Gaze [242] shows relatively low error rates, notably achieving the lowest error in the $\mathcal{D}RT \rightarrow \mathcal{D}GC$ task with an error of 13.20°. However, MTGH-Net reports the lowest angular errors of 8.29° and 11.41° in the $\mathcal{D}RT \rightarrow \mathcal{D}MP$ and $\mathcal{D}RT \rightarrow \mathcal{D}GC$ tasks respectively, outperforming Dilated-Net by approximately 8% and surpassing Eth-Gaze by about 14%.

In conclusion, MTGH-Net outperforms the exiting state of the art methods on the four cross dataset evaluation tasks with up to 27% improvement in gaze reliability. This shows its robust capability to accurately generalize across different domains. This proves that MTGH-Net is able to effectively transfer knowledge across different domains and improve gaze estimation performance in unseen datasets. This success can be attributed to the strong feature representation and enhanced knowledge sharing achieved through the multitask approach.

4.5.2.2 Ablation Study

A comprehensive ablation study is conducted to show the contributions of different components of the MTGH-Net to its overall performance. Several experiments are conducted to further test the robustness of the proposed MTGH-Net, including the ablation study on the multitask approach, the proposed gaze representation and loss

Methods	$\boldsymbol{\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}}$	$\boldsymbol{\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}}$	$\boldsymbol{\mathcal{D}_{RT} \rightarrow \mathcal{D}_{MP}}$	$\boldsymbol{\mathcal{D}_{RT} \rightarrow \mathcal{D}_{GC}}$
STG-Net	7.60°	17.95°	8.75°	12.72°
MTGH-Net	**6.0°**	**15.50°**	**8.29°**	**11.41°**

Table 4.3: Gaze performance comparison between STG-Net (single task gaze model) and the proposed MTGH-Net.

Methods	$\boldsymbol{\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}}$	$\boldsymbol{\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}}$	$\boldsymbol{\mathcal{D}_{RT} \rightarrow \mathcal{D}_{MP}}$	$\boldsymbol{\mathcal{D}_{RT} \rightarrow \mathcal{D}_{GC}}$
MTGH-Net-SL2	7.26°	15.78°	8.31°	11.90 °
MTGH-Net	**6.0°**	**15.50°**	**8.29°**	**11.41°**

Table 4.4: Gaze performance comparison between MTGH-Net-SL2 (network with spherical angles and L2 loss) and the proposed MTGH-Net.

function, the network architecture and the backbones. In all experiments, rotation representation and geodesic loss are adapted for the head pose task.

1) Contribution of the multitask approach: To validate the impact of the multitask approach on the overall performance of MTGH-Net, a single task gaze network architecture (STG-Net) is compared with MTGH-Net. The STG-Net network utilizes a ResNet-50 backbone and a final fully connected layer to predict solely the gaze rotation matrix. To ensure a fair comparison, STG-Net utilizes the same training configurations as those used in MTGH-Net.

Table 4.3 shows the comparative performance of the four cross dataset evaluation tasks. The results highlight the superiority of MTGH-Net in all evaluation tasks compared to STG-Net. It can be observed that MTGH-Net exhibits a significant improvement over the STG-Net, enhancing cross-dataset performance by 21%, 13.6%, 6%, and 10.2% for transitions from ($\mathcal{D}_G$) to ($\mathcal{D}_M$), ($\mathcal{D}_G$) to ($\mathcal{D}_C$), ($\mathcal{D}_R$) to ($\mathcal{D}_M$), and ($\mathcal{D}_R$) to ($\mathcal{D}_C$), respectively. These results confirm that MTGH-Net leverages the synergistic correlation between gaze and head pose estimation tasks, benefiting from the shared features and diverse data encompassing both tasks. This approach enhances its generalizability across different datasets, underscoring the advantage of a multitask learning framework in gaze estimation endeavors.

2) Contribution of gaze rotation representation and loss function: To evaluate the importance of employing a rotation matrix and geodesic loss on the overall gaze performance, a multitask network, MTGH-Net-SL2, is proposed and compared with MTGH-Net. MTGH-Net-SL2 utilizes spherical angles and L2 loss for its gaze estimation component, while MTGH-Net incorporates a rotation matrix and geodesic loss for this part. To ensure comparability, both networks are subjected to the same

Methods	$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}$	$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}$	$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{MP}$	$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{GC}$
MTGH-Net-V	6.86°	16.62°	9.25°	11.50 °
MTGH-Net	**6.00°**	**15.50°**	**8.29°**	**11.41°**

Table 4.5: Gaze performance comparison between standard multi-task approach (MTGH-Net-S) and the proposed MTGH-Net.

training configurations.

The results, shown in Table 4.4, clearly show the superior gaze generalization performance of MTGH-Net over the MTGH-Net-SL2. This marked superiority validates the hypothesis that a unified representation for both gaze and head pose tasks, facilitated by the rotation matrix and geodesic loss, enhances training efficiency. Moreover, it addresses the critical issues associated with discontinuity inherent in Euler and spherical angle representation, ensuring a more stable and reliable learning process.

3) Contribution of network architecture: A third experiment is conducted to assess the effect of the modified ResNet-50 architecture on the overall gaze accuracy. MTGH-Net hypnosis that the integration of the last residual block of ResNet50 and global average pooling layer into task-specific top layers for each task help enable the extraction of distinct high-level features for gaze and head pose estimation tasks, thus potentially enhancing their performance. To empirically validate this premise, a comparative framework, MTGH-Net-V is proposed and compared with MTGH-Net model. MTGH-Net-S framework employs a conventional ResNet-50 architecture followed by two fully connected layers, one each for gaze and head pose estimation tasks. Similarly to MTGH-Net, MTGH-Net-V utilizes a rotation representation and geodesic loss for the gaze estimation task, adhering to the same training setup to ensure fair comparison.

The comparative results, detailed in Table 4.5, highlight the superior gaze generalization performance of MTGH-Net over the MTGH-Net-V. Specifically, these results validate the architectural strategy, demonstrating that integrating the last residual block and average pooling layer for task-specific layers significantly exceeds the performance achievable with a baseline ResNet-50 architecture. This evidence confirms the importance of the architectural approach MTGH-Net in generating rich and high-level features that directly contribute to improving the accuracy of gaze estimation.

4) Choice of backbone: A final experiment to examine the impact of model size on the performance of the MTGH-Net by swapping the use of a ResNet-50 backbone with a more compact ResNet-18 backbone. Although the reliability of MTGH-Net with ResNet-50 has already shown its superiority detailed in Tables 4.1 and 4.2, this experiment aims to show the impact of model size on the performance of MTGH-

Methods	$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}$	$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}$	$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{MP}$	$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{GC}$
ResNet-18	6.04°	15.71°	8.54°	11.45°
ResNet-50	**6.0°**	**15.50°**	**8.29°**	**11.41 °**

Table 4.6: Gaze performance of the proposed MTGH-Net adapting different backbones of ResNet-18 and ResNet-50.

Table 4.7: Comparisons with the SOTA methods on the two cross-dataset head pose tasks. MTGH surpasses SOTA methods on the BIWI dataset and achieves comparable results to the SOTA method on the AFLW2000 dataset.

Methods	$\mathcal{D}_{WL} \rightarrow \mathcal{D}_{AF}$				$\mathcal{D}_{WL} \rightarrow \mathcal{D}_{BI}$			
	Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
HopeNet ($\alpha = 2$) [158]	6.47	6.56	5.44	6.16	5.17	6.98	3.39	5.18
HopeNet ($\alpha = 1$) [158]	6.92	6.64	5.67	6.41	4.81	6.61	3.27	4.90
FSA-Net [225]	4.50	6.08	4.64	5.07	4.27	4.96	2.76	4.00
HPE [81]	4.80	6.18	4.87	5.28	3.12	5.18	4.57	4.29
QuatNet [75]	3.97	5.62	3.92	4.50	**2.94**	5.49	4.01	4.15
Viet et al [189]	-	-	-	-	4.62	4.29	4.52	4.48
WHENet-V [264]	4.44	5.75	4.31	4.83	3.60	4.10	2.73	**3.48**
WHENet [264]	5.11	6.24	4.92	5.42	3.99	4.39	3.06	3.81
TriNet [22]	4.04	5.77	4.20	4.67	4.11	4.76	3.05	3.97
FDN [236]	**3.78**	5.61	3.88	4.42	4.52	4.70	**2.56**	3.93
MTGH-Net	4.46	**5.11**	**3.54**	**4.37**	4.65	**3.77**	2.74	3.72

Net. The results presented in Table 4.6 indicate that replacing ResNet-50 with the smaller ResNet-18 backbone incurs only a marginal performance drop of max 0.2°in all cross-dataset evaluation tasks. Remarkably, this minor drop does not detract from the ability of MTGH-Net to retain state-of-the-art performance. This underscores the robustness of the MTGH-Net and its ability to deliver top gaze estimation accuracy independent of the backbone's complexity. This attribute of model flexibility is particularly significant for deployment in real-world scenarios, where computational resources may be constrained. The ability of MTGH-Net to preserve its accuracy while accommodating smaller, more efficient backbones ensures its applicability across a wide spectrum of devices and environments, potentially enhancing the feasibility and reach of advanced gaze estimation solutions.

4.5.2.3 Head pose cross dataset evaluation

The MTGH-Net is benchmarked against state-of-the-art head pose estimation methods. All the methods considered for comparison utilize the backbones to predict solely head pose as a single task. The evaluation includes widely used datasets for head pose, including AFLW2000 and BIWI datasets. Table 4.7 presents the MAE of MTGH-Net and the state of the art head pose estimation methods on the two cross dataset evaluation tasks include $\mathcal{D}_{WL} \rightarrow \mathcal{D}_{AF}$ and $\mathcal{D}_{WL} \rightarrow \mathcal{D}_{BI}$.

For $\mathcal{D}_{WL} \rightarrow \mathcal{D}_{AF}$ task, the proposed MTGH-Net achieves the best overall MAE of 4.37 on this task, showing effective overall reliability compared with other methods. Besides the overall error, MTGH-Net has the lowest pitch and yaw errors of 5.11° and 5.11°, respectively and a competitive result on the yaw angle. For $\mathcal{D}_{WL} \rightarrow \mathcal{D}_{BI}$ task, MTGH-Net achieves a competitive results on the overall MAE reporting the second lowest error of 3.72, ranking it among the top performing methods. Further, the proposed MTGH-Net achieves the lowest error rate on the pitch angel of 3.77, showing it is ability to maintain high performance across domains.

The results underscore the superiority of MTGH-Net over competing state of the art approaches on the AFLW2000 dataset, while providing comparable results on the BIWI dataset. This quantitative analysis confirms that the multitask approach successfully maintains performance in head pose prediction with the same number of parameters as a single-task approach while achieving SOTA results on one of the test datasets. The integration of head pose and gaze estimation within a single network not only enhances accuracy but also significantly reduces computational costs by approximately 50% compared to estimating them independently. This shows the efficacy of leveraging the synergy between head pose and gaze estimation in a multitask framework, offering significant computational and performance advantages.

4.5.3 Qualitative Results

The MTGH-Net has demonstrated its effectiveness through quantitative results, setting new benchmarks in generalization performance across various datasets. However, the qualitative analysis provides invaluable insights into the model's practical capabilities through examples and visualizations. This section investigates the visual aspects of MTGH-Net illustrating its effectiveness through various examples and visualizations.

4.5.3.1 Visualization of MTGH-Net predictions

The predictions of MTGH-Net are visualized across a range of frames from videos of the 300LW dataset [164] in Figure 4.6. These frames are carefully chosen to evaluate challenging scenarios, including a wide range of subjects, environments, lighting conditions, and camera perspectives, with particular attention to situations where the gaze direction significantly differs from the head pose (eye head interplay). The illustrations in the figure highlight the ability of MTGH-Net to accurately estimate both gaze and head pose even in these challenging conditions. By demonstrating robust performance across diverse conditions, MTGH-Net validates its utility and reliability for real-world

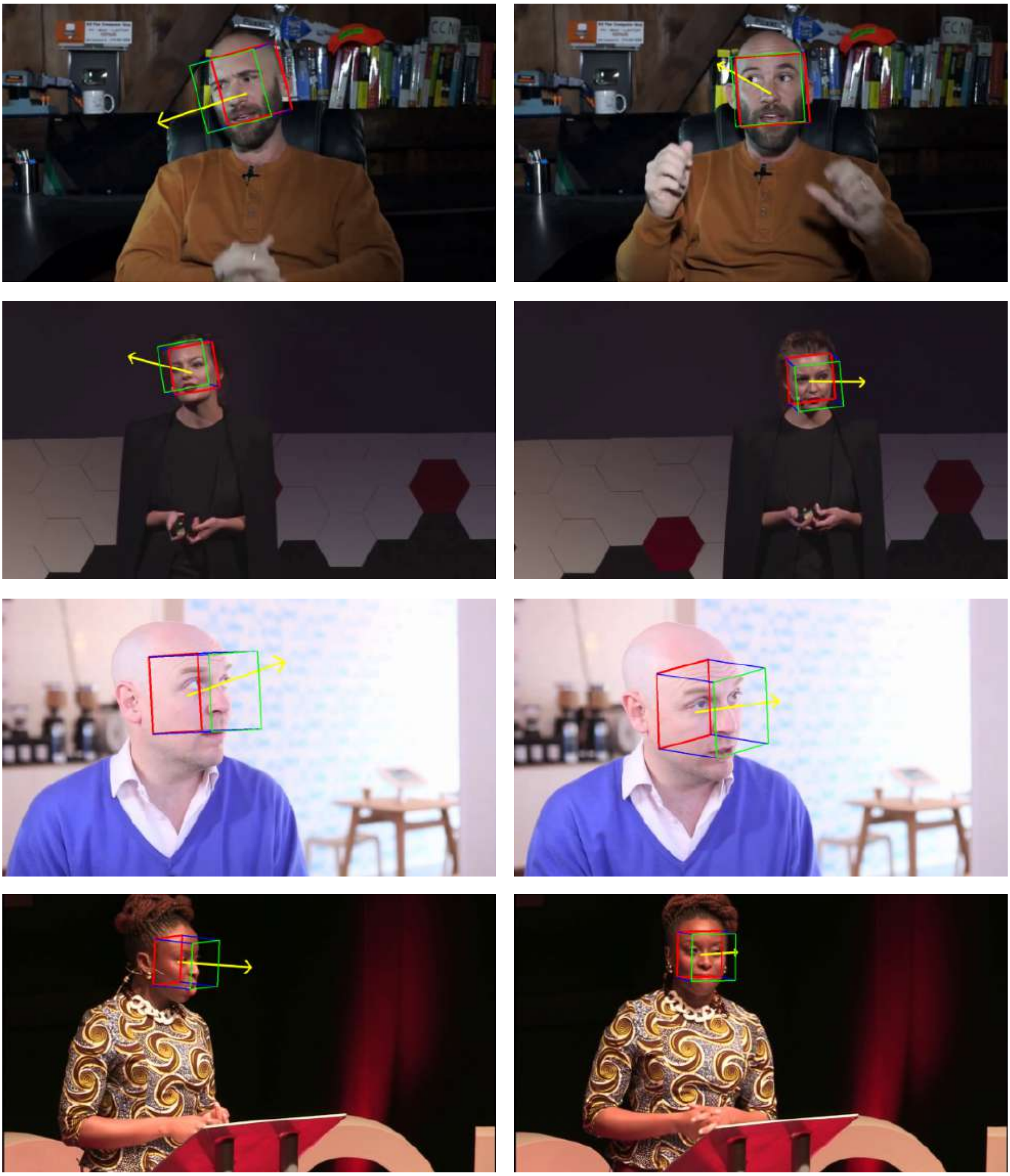


Figure 4.6: MTGH-Net simultaneous gaze and head pose predictions of some frames extracted from videos in the 300VW [164] dataset. The red arrow represents gaze, and the wire frame cube represents head pose.

applications, showcasing its potential to generalize well beyond the training data and adapt to the natural variations encountered in practical settings.

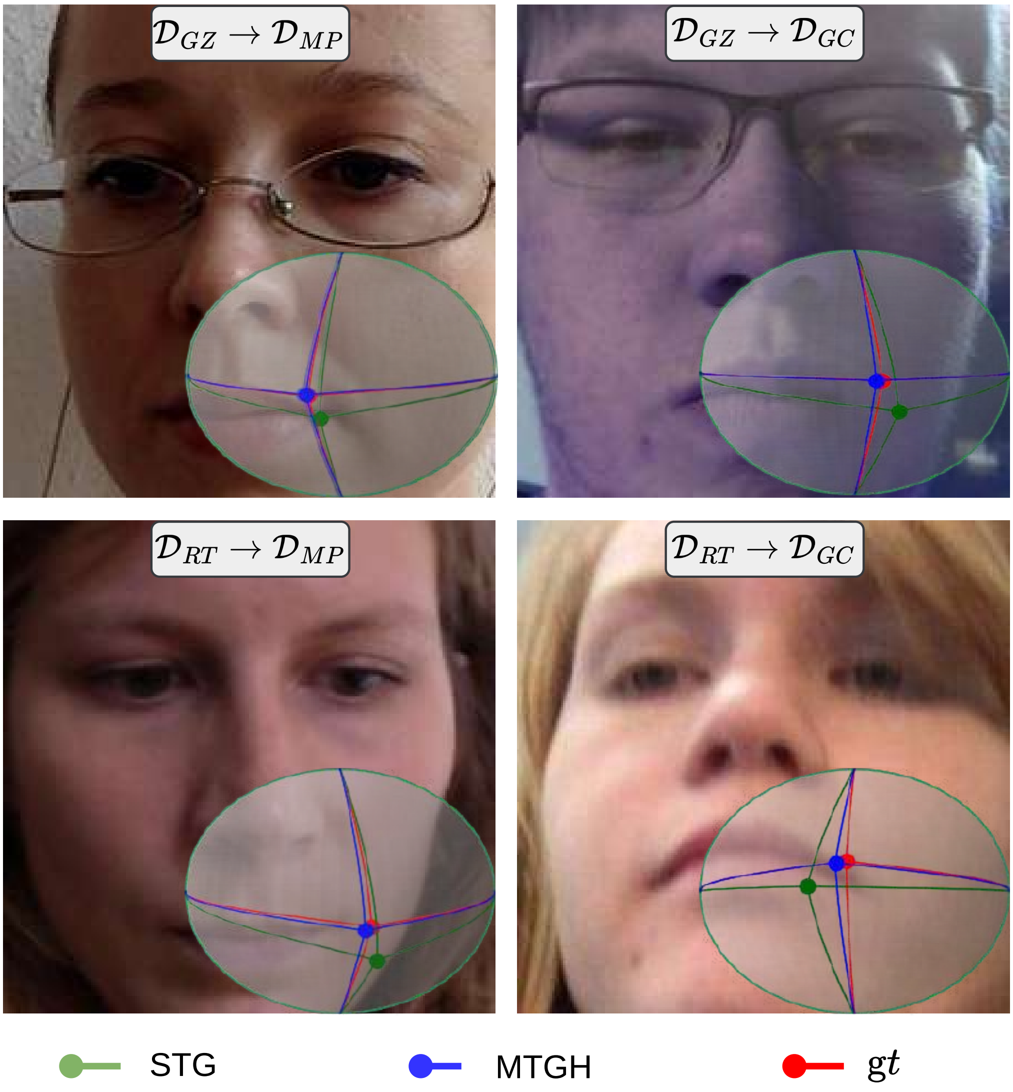

$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}$
$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}$
$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{MP}$
$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{GC}$
STG
MTGH
gt

Figure 4.7: Visual results example of estimated 3D gaze. Red points represent the ground-truth gaze direction, green and blue points represent the predictions of the single task gaze model and the proposed MTGH-Net, respectively.

4.5.3.2 MTGH-Net Comprehension

The results in Table 4.3 have demonstrated the superiority of MTGH-Net over STG-Net in gaze generalization performance. To substantiate these results, a visualization of MTGH-Net, STG-Net, and ground truth gaze values is presented in Figure 4.7. This visualization encompasses four gaze cross-dataset tasks: $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}$, $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}$, $\mathcal{D}_{RT} \rightarrow \mathcal{D}_{MP}$, and $\mathcal{D}_{RT} \rightarrow \mathcal{D}_{GC}$. Observations reveal that MTGH-Net's predictions are closely aligned with ground truth values, whereas STG-Net's predictions deviate significantly from the ground truth in these tasks.

This clear alignment emphasizes the enhanced capability of MTGH-Net to adapt to new and diverse domains, surpassing that of the STG model. Such adaptability is attributed to MTGH-Net's ability to learn rich feature representations through a mul-

Table 4.8: Performance comparison with SOTA domain adaptation methods. Although MTGH don't use any data from the target domain, It provides SOTA results among domain adaption methods.

Method	**Target Samples**	$\boldsymbol{\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}}$	$\boldsymbol{\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}}$
GazeAdv [198]	> 100	8.19°	19.21°
Gaze360Adv [95]	> 100	7.45°	17.12°
DAGEN [64]	~ 500	6.61°	-
ADDA [181]	~ 500	8.59°	-
UMA [19]	~ 100	9.17°	-
RUDA [11]	< 100	6.20°	-
PNP-GA [116]	< 100	6.18°	-
MTGH-Net	N/A	**6.04°**	**15.71°**

titask approach, contrasting with the narrower focus of STG-Net in gaze dataset only. By effectively leveraging a multitask framework, MTGH-Net demonstrates robust feature extraction capabilities that generalize better than those of STG-Net. This shows how it facilitates the distillation of knowledge from various domains under different constraints, thereby enhancing the accuracy of gaze and head pose predictions.

4.5.4 Comprehensive Analysis

This section delves into the nuanced performance of the model, offering a deeper understanding of its capabilities through performance analysis against domain adaptation and degree wise error analysis.

4.5.4.1 Performance analysis against domain adaptation

To show the robustness of MTGH-Net, benchmarks are performed against state-of-the-art gaze domain adaptation methods on the two cross dataset tasks: $\boldsymbol{\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}}$ and $\boldsymbol{\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}}$. These methods include GazeAdv [198], Gaze360Adv [95], DAGEN [64], RUDA [11] and PNP-GA [116]. Additionally, it is compared with methods include ADDA [181] which is originally proposed for classification problems, and UMA [19] which is designed for hand segmentation tasks. Performance results for these competing methods are directly reported from [11, 116], which they implemented all these methods using ResNet-18 backbone. For a fair comparison, a switching from a ResNet-50 to a ResNet18 backbone is performed, aligning with the backbone used by the mentioned domain adaptation methods. Despite these modifications, it is acknowledged that directly comparing the MTGH-Net with domain-adaptation approaches might not be entirely fair. This discrepancy arises from the fundamental difference in methodologies, where domain adaptation methods typically utilize target

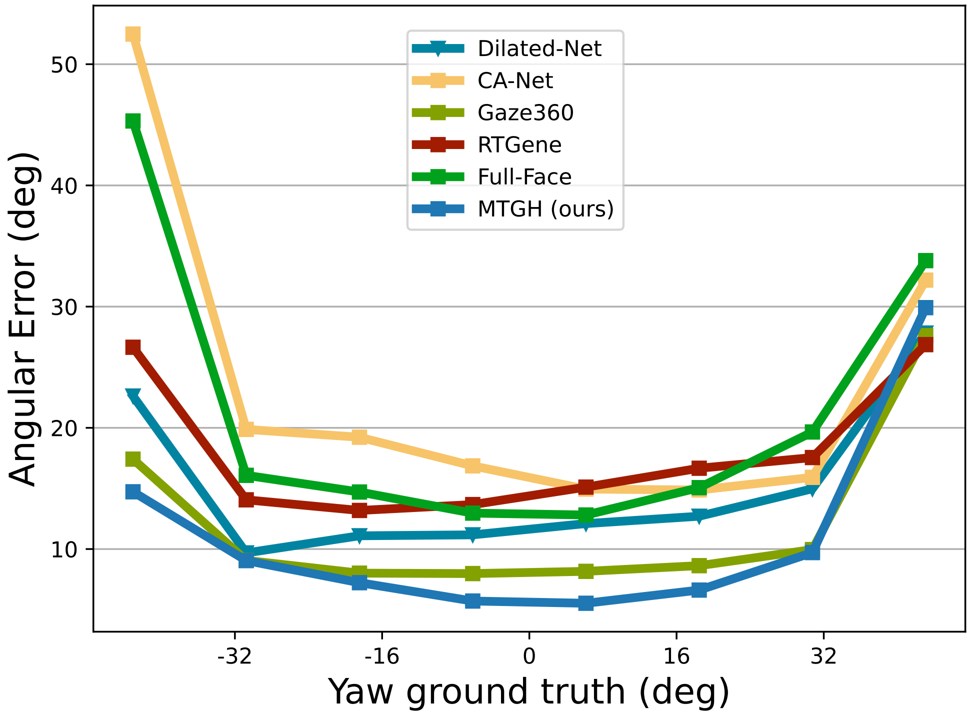


Figure 4.8: Degree-wise error analysis of MTGH-Net and the state of the art methods on the yaw angel of the $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}$ task.

domain samples during their training process, unlike the MTGH-Net.

The results presented in the table 4.8 illustrate that the MTGH-Net outperforms all the domain adaptation methods mentioned in tasks of $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}$ and $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}$, thereby establishing new SOTA benchmarks. This superior performance underscores the robustness of the MTGH-Net and its ability to learn a singular optimal model capable of generalizing well across multiple domains. In contrast, domain-adaptation methods typically require learning a distinct model for each target domain, complicating their applicability in real-world scenarios. This distinction highlights the potential of the MTGH-Net in providing a more universally applicable solution for gaze estimation tasks, further demonstrating its relevance and utility in advancing the field.

4.5.4.2 Degree-wise error analysis

To comprehensively assess MTGH-Net, a degree-wise error analysis is conducted focusing on $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}$ task. This task asses the generalization from the Gaze360 dataset ($\mathcal{D}_{GZ}$) to the MPIIFaceGaze dataset ($\mathcal{D}_{MP}$). setting the MTGH-Net against a state-of-the-art methods, including Full-Face [251], RT-Gene [54], CA-Net [30], Dilated-Net [27], and Gaze360 [95]. The MPIIFaceGaze dataset, known for its varied pitch and yaw angle ranges, served as the benchmark for this comparison. To accommodate the dataset's variability, errors in pitch angles are categorized into intervals of 10°, and yaw angles into intervals of 20°, as depicted in Fig. 4.8 and Fig. 4.9.

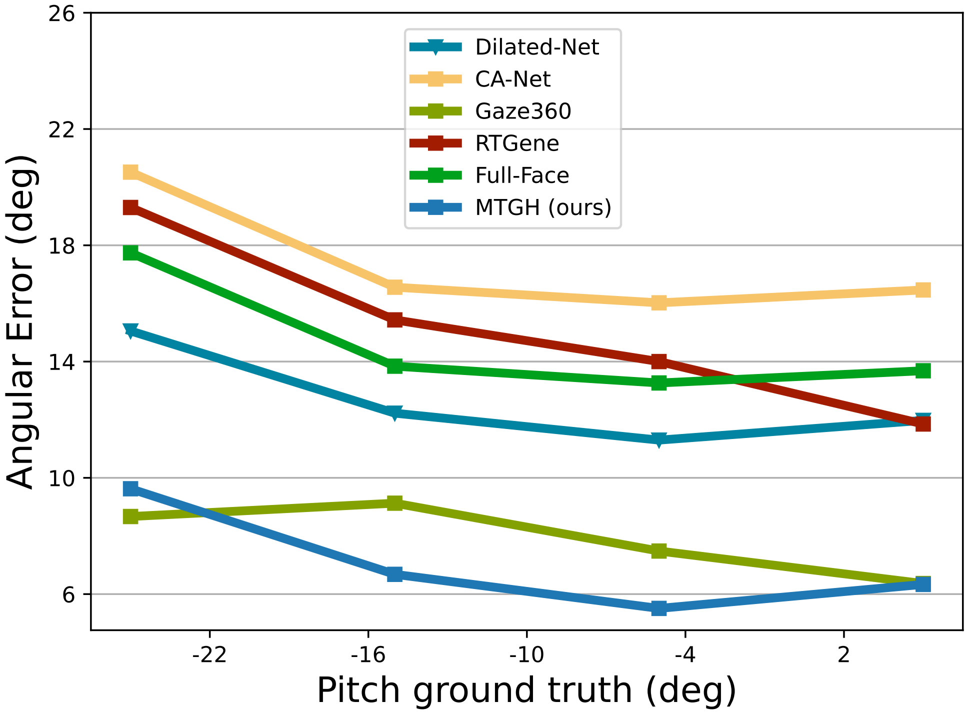


Figure 4.9: Degree-wise error analysis of MTGH-Net and the state of the art methods on the pitch angel of the $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}$ task

The analysis revealed that the MTGH-Net consistently exhibited the lowest error rates across almost all intervals when compared with the SOTA methods. This result show the superior reliability of MTGH-Net in predicting accurate gaze direction when generalizing to new un seen domains. These results are indicative of the advanced capability of the MTGH-Net in handling the complexities associated with varying gaze angles, further emphasizing its potential in significantly improving gaze estimation tasks across diverse environments and constraints.

4.6 Conclusion

This chapter introduces MTGH-Net, an innovative model designed to leverage the intrinsic correlation between gaze and head pose through a supervised multi-task approach. It employs an advanced training strategy that integrates two distinct datasets, each annotated for gaze and head pose. This method effectively minimizes the impact of variations in appearance and head pose, enriching the model's ability to generalize gaze estimation across diverse settings by distilling knowledge from images under varied conditions.

To avoid task bias, MTGH-Net utilizes a rotation matrix formalism for gaze annotation and a continuous 6D rotation matrix representation for direct regression of both gaze and head pose tasks. This approach is further enhanced by a geodesic distance-based loss function, which penalizes the network to ensure that the model accurately reflects the complexities of spatial orientation.

The framework has undergone rigorous evaluation against public datasets for gaze and head pose estimation, demonstrating superior efficiency and robustness, and markedly outperforming state-of-the-art methods. Notably, it achieves up to a 21% improvement in gaze generalization performance over single-task approaches. Moreover, MTGH-Net surpasses existing state-of-the-art domain adaptation and gaze generalization techniques in various cross-dataset gaze tasks, highlighting its effectiveness in leveraging multi-task learning to enhance performance.

Furthermore, the integration of the head pose task enhances gaze generalization performance. The strong correlation between the two tasks and the valuable information shared through the multi-task approach improve regularization during training, leading to top performance for both gaze and head pose in the most popular benchmarks. This integration allows for the accurate prediction of both gaze and head pose within a single network, significantly reducing the computational cost of the model by approximately 50% compared to estimating them independently.

5 Improving Gaze Estimation Efficiency Using Mobile Network And Progressive Attention Mechanisms

5.1 Introduction

Gaze estimation plays a crucial role in the field of computer vision. It supports a variety of applications that range from improving human computer interaction to advances in augmented reality [39, 150], autonomous driving [1, 59, 70, 79, 139, 190] and human robot interaction (HRI) [2, 71, 99, 100, 169]. Traditionally, the field of gaze estimation relied on model-based approaches [68, 188, 213] that required specialized hardware. These complex requirements drove up costs and limited the scalability of gaze estimation solutions. Appearance-based methods [9, 151, 170] take advantage of inexpensive and available cameras and replace hardware based approaches that guarantee scalable and flexible solutions to improve gaze estimation. In addition, the advent of deep learning [146, 248], particularly through advances in convolutional neural networks (CNNs), led to a new era for the advancement of gaze estimation methods.

Deep learning gaze estimation methods have changed from using the high computational approach of the eye face [27, 36] to adapting the efficient face-only approach [136, 250, 250] to simplify the gaze estimation process. The main issue of this approach is that the eyeball comprises only a small part of the face image, which makes it difficult to extract fine-grained gaze information. Furthermore, traditional CNNs often fail to discriminate between gaze-relevant features and other facial characteristics because of their tendency to focus on larger and more continuous regions of the image. To address this, methods adapt complex CNNs and transformer models [32, 135] to extract rich information from eye region to increase accuracy. This models inherently demand substantial computational resources. The challenge is further increased in multimodal gaze estimation scenarios, where additional inputs such as head pose is necessary for specific applications. These multimodal scenarios are increasingly common in applications such as HRI, augmented reality, and autonomous driving, each adding complexity and computational load which pose significant challenges for achieving real-time performance.

In this chapter, MGAZE-Net is proposed as an efficient solution to provide an optimal balance between gaze performance and computational cost. MGAZE-Net is able to capture fine-grained gaze features from facial images while using a lower computational cost. First, MGAZE-Net leverages an efficient architecture utilizing inverted residuals and linear bottlenecks introduced in MobilNetV2 [161, 200, 240] to solve the computational cost problem. Then, it is characterized by the integration of a progressive combination of attention mechanisms, including Squeeze-and-Excitation (SE), Convolutional Block Attention Module (CBAM), and Coordinate Attention (CA). Each of these attention layers has the advantage of emphasizing crucial gaze information within the face image. Finlay, Global depth-wise Convolution (GDConv) is used in the last stage of MGAZE-Net to capture discriminative information from diverse units within the final feature map. This strategic fusion not only highlights critical eye features, but also enables the model to effectively capture both local and global spatial relationships. Unlike traditional methods, MGAZE-Net achieves this without additional computational overhead which make it ideal for use in mobile and embedded systems. In addition, MGAZE-Net uses a rotation matrix formalism with geodesic-based loss function for the gaze representation. This approach mitigates discontinuity and ambiguity issues and provides a continuous and unambiguous representation that improves the learning efficiency of the model and the accuracy of gaze estimation. MGAZE-Net demonstrates state of the art balance between performance and computational cost on four popular datasets.

5.2 Key Contributions

This chapter introduces a novel network that significantly advances the state-of-the-art of gaze estimation by addressing the trade-off between accuracy, reliability and efficiency. This novel method can capture fine-grained features from the eye region within facial images without adding additional computational cost. The key contributions of this chapter are presented in detail:

- MGAZE-Net: A novel lightweight CNN based on mobile network architecture augmented with a progressive combination of attention mechanisms. It use MobilNetV2 bottleneck integrated with combination of attention mechanisms include SE, CBAM and CA modules. This hierarchical integration of attention mechanisms is designed to systematically emphasize important gaze information by capturing both local and global spatial relationships within facial images. The strategic placement of these mechanisms allows MGAZE-Net to extract fine-grained features relevant to gaze estimation with remarkable efficiency, without the computational overhead typically associated with deep CNN models and transformers.

- Rotation Matrix Formalism for Gaze Representation: MGAZE-Net uses a rotation matrix formalism to represent gaze direction. This novel approach mitigates the problems of discontinuity and ambiguity with spherical angle representation

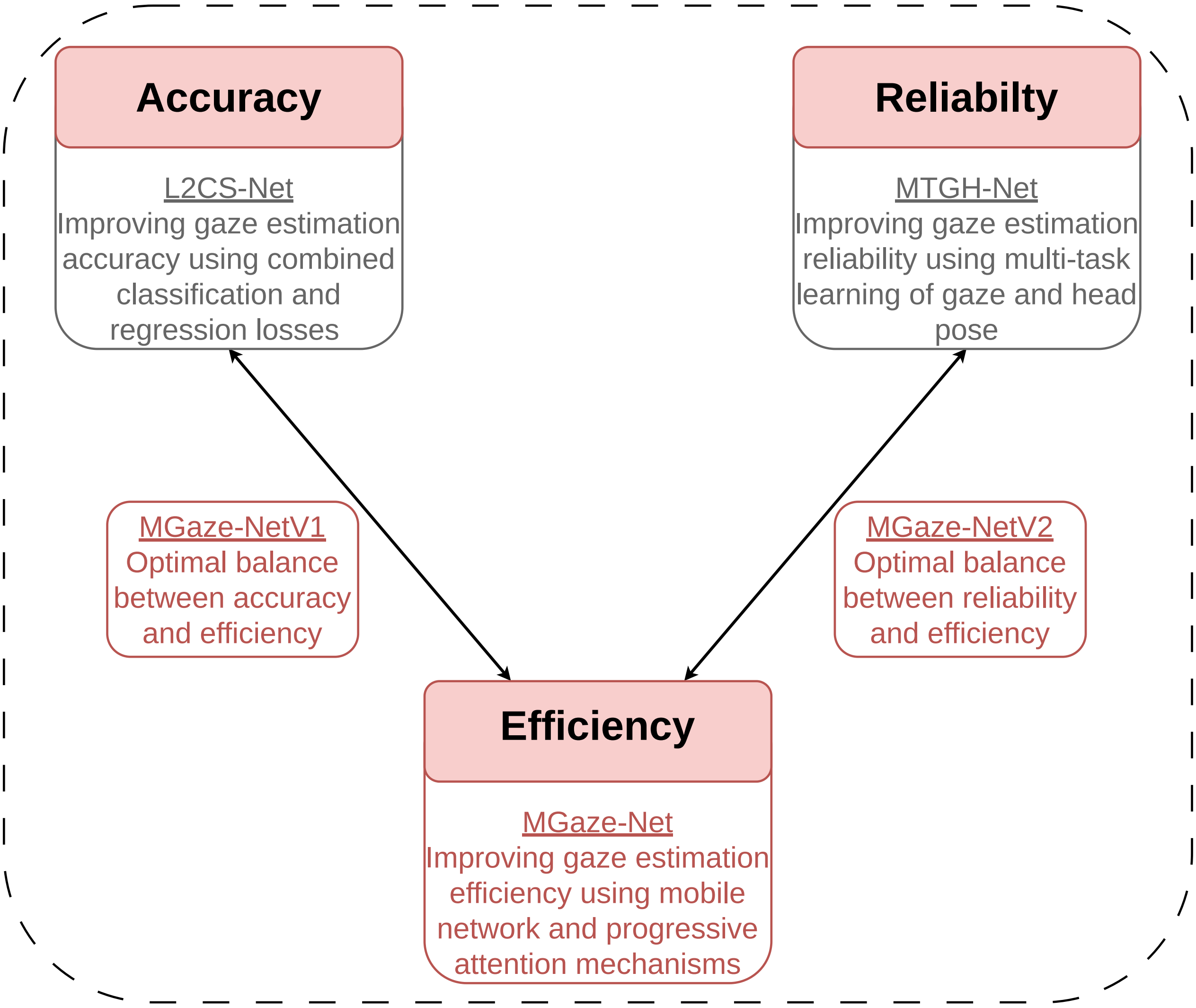

Accuracy
L2CS-Net
Improving gaze estimation accuracy using combined classification and regression losses
Reliabilty
MTGH-Net
Improving gaze estimation reliability using multi-task learning of gaze and head pose
MGaze-NetV1
Optimal balance between accuracy and efficiency
MGaze-NetV2
Optimal balance between reliability and efficiency
Efficiency
MGaze-Net
Improving gaze estimation efficiency using mobile network and progressive attention mechanisms

Figure 5.1: The goal of this chapter is to design an accurate gaze estimation model that can extract rich gaze features from the small eye region in the facial images.

and provides a continuous, unambiguous and geometrically accurate method for gaze representation.

- Geodesic-based Loss Function for Enhanced Training Efficiency: MGAZE-Net introduces a geodesic loss function which utilizes the geometric properties of rotation matrices and provides a more accurate measure of the discrepancy between predicted and ground truth gaze values. It facilitates more efficient training by directly penalizing errors in the rotation space, significantly improving the performance MGAZE-Net.

- Extensive Validation on Challenging Datasets: The efficacy and robustness of the proposed MGAZE-Net are rigorously validated through extensive experiments on four challenging datasets: MPIIFaceGaze, GazeCapture Gaze360, and RT-GENE. The comprehensive evaluation demonstrates that MGAZE-Net succeed to improve the balance between performance and computational cost.

- Ablation Study: Through detailed ablation study, the contributions of each component of the proposed MGAZE-Net is analysed. These studies reveal the critical role of network architecture and gaze representation.

The remainder of this chapter is organized as follows. Section 5.3 presents the relevant prior works. Section 5.4 presents the details of the proposed MGAZE-Net framework. In Section 5.5, experimental, quantitative and quantitative analysis of the MGAZE-Net is presented. Finally, Section 5.6 contains the conclusion.

5.3 Prior Work

5.3.1 Gaze estimation

The field of appearance-based gaze estimation has experienced significant advances in methodologies and techniques aimed at accurately mapping image intensities to human gaze directions. Initially, efforts concentrated on learning person-specific mapping functions, including linear interpolation [173], adaptive linear regression [122], and Gaussian process regression [208]. These methods achieved reasonable accuracy within constrained environments, such as subject-specific settings and fixed illuminations, but their performance significantly declined in unconstrained scenarios.

A pivotal shift occurred with the emergence of CNN-based gaze estimation methods, which aimed to capture intricate non-linear mappings between images and gaze directions. Zhang et al.[253] introduced a simple VGG network architecture for gaze prediction using single-eye images. A CNN spatial weighting scheme was devised to emphasize critical gaze information in faces[251]. Multichannel networks were proposed by Krafka et al.[107], leveraging eye images, full-face images, and face grid data to enhance the extraction of essential gaze information. Chen et al.[27] utilized dilated convolutions to preserve high-level features while retaining spatial resolution.

Further enhancements included the integration of head pose vector and ensemble techniques with VGG-based gaze prediction by Fischer et al.[54], which led to increased accuracy. Cheng et al.[36] introduced FAR-Net, a method to estimate 3D gaze angles using asymmetric weighted loss functions for each eye. A novel hybrid model, ID-ResNet, incorporated a residual neural network structure and a layer of personal identity to accommodate geometric variations between subjects [201]. MeNets [219] fused statistical models with deep learning, introducing mixed-effect models within CNN architectures. Wang et al.[199] combined adversarial learning with Bayesian approaches, resulting in improved gaze generalization performance. CA-Net[30] allowed for the prediction of primary gaze angles from face images and their refinement using eye crop residuals.

Kellnhofer et al.[95] leveraged Long Short-Term Memory (LSTM) networks and sequences of face images to predict gaze angles, introducing a pinball loss to jointly regress gaze directions and error bounds. Murthy et al.[14] proposed parallel networks for each image of the eye, using convolution and attention-based networks to refine gaze prediction.

Recent advancements have seen the integration of transformer-based models to extract gaze features from images with high variance [32, 135]. These models effectively filtered irrelevant information using convolutional projection and maintained detailed image features through convolution layers. While transformers achieved notable gaze performance, they also introduced a trade-off of increased computational cost compared to conventional CNN-based models.

5.3.2 Attention mechanisms

The ability of humans to effectively identify significant areas within complex visual scenes has inspired researchers to introduce attention mechanisms into computer vision systems [174]. Attention mechanisms have proven valuable in various computer vision tasks, including image classification [130, 226] and image segmentation [138, 227, 238]. Various attention modules have been developed to improve the performance of computer vision tasks, including channel attention and spatial attention [63]. The most effective attention mechanisms used in different computer vision tasks include SE [77], CBAM [209], and CA [72]. These mechanisms have been widely used in recent mobile networks [26, 77] and have been shown to be key components in achieving state-of-the-art performance.

5.4 Framework

MGAZE-Net integrates a progressive combination of attention mechanisms within a lightweight CNN architecture. This unique configuration is tailored to extract fine-grained features from the eye region with efficient computational cost. A complete pipeline of the proposed framework is depicted in Figure 4.3.

5.4.1 Preliminaries

This section introduces fundamental concepts and methodologies that underpin the proposed approach in this chapter. The discussion begins with an exploration of depth-wise separable convolutions, a powerful method used to enhance the efficiency and performance of CNNs. It then delves into the MobileNet architecture, which leverages depth-wise separable convolutions to achieve high efficiency in neural network models that make them suitable for mobile and embedded applications. Finally, the section explores various attention mechanisms that further enhance the performance of these models by enabling them to focus on the most informative parts of the input data.

5.4.1.1 Depth-wise Separable Convolution

Depth-wise separable convolution applies a single filter to each input channel independently, performing spatial filtering without altering the depth. This step is responsible for extracting spatial features from each channel without combining information across channels. The depth-wise separable convolution significantly reduces the number of parameters and the computational complexity compared to the standard convolution layer. Specifically, for a given layer with an input size of $D_F \times D_F$ and M channels, applying a standard convolution with a kernel size of $D_K \times D_K$ and N output channels is computed using:

$$G_{k,l,n} = \sum K_{i,j,m,n} \cdot F_{k+i-1,l+j-1,m} \tag{5.1}$$

$G_{k,l,n}$ represents the output feature map, where k and l are spatial indices and n is the index for the output channel. $K_{i,j,m,n}$ denotes the kernel weights, with i and j being the spatial indices within the kernel, m representing the input channel and n the output channel. $F_{k+i-1,l+j-1,m}$ corresponds to the input feature map, with $k+i-1$ and $l+j-1$ adjusting the spatial location according to the size of the kernel, and m indicating the input channel.

Standard convolution has a computational cost of:

$$D_K \times D_K \times M \times N \times D_F \times D_F \tag{5.2}$$

depth-wise separable convolution operates differently from standard convolution by applying a single filter per input channel (input depth), effectively separating the filtering process across channels. The process is broken down into two distinct layers: depth-wise convolution and pointwise convolution. The computational formula for depth-wise separable convolution can be written as

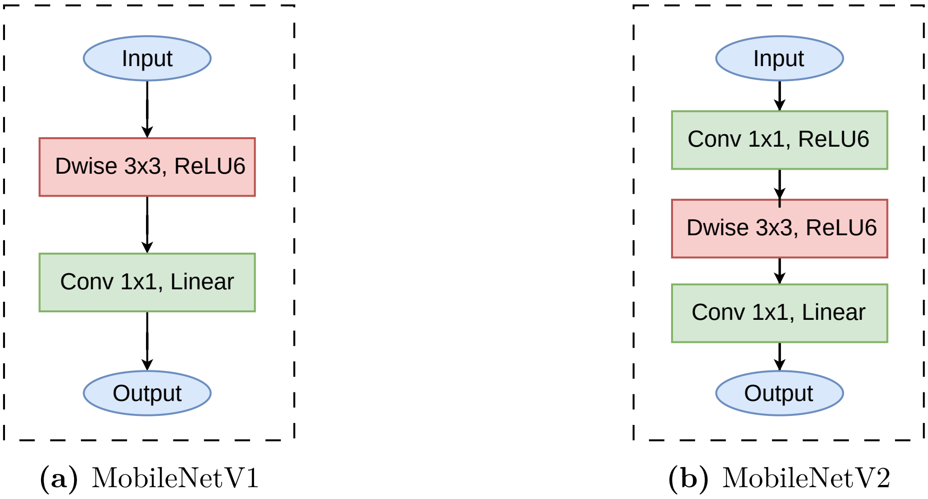
$$\hat{G}_{k,l,m} = \sum_{i,j} \hat{K}_{i,j,m} \cdot F_{k+i-1,l+j-1,m} \tag{5.3}$$

In this equation, $\hat{G}_{k,l,m}$ represents the output of the depth-wise convolution for each channel, where k and l are spatial indices, and m corresponds to the specific input channel being processed. $\hat{K}_{i,j,m}$ denotes the kernel weights for the depth-wise convolution, with i and j indicating the spatial positions within the kernel, and m specifying the channel to which the kernel is applied. The input feature map is represented by $F_{k+i-1,l+j-1,m}$, with spatial indices $k+i-1$ and $l+j-1$ adjusted for the position of the kernel, and m indicating the input channel.

Depth-wise convolution have a computational cost of:

$$D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F \tag{5.4}$$

depth-wise separable convolution significantly reduces computational complexity by



(a) MobileNetV1 **(b)** MobileNetV2

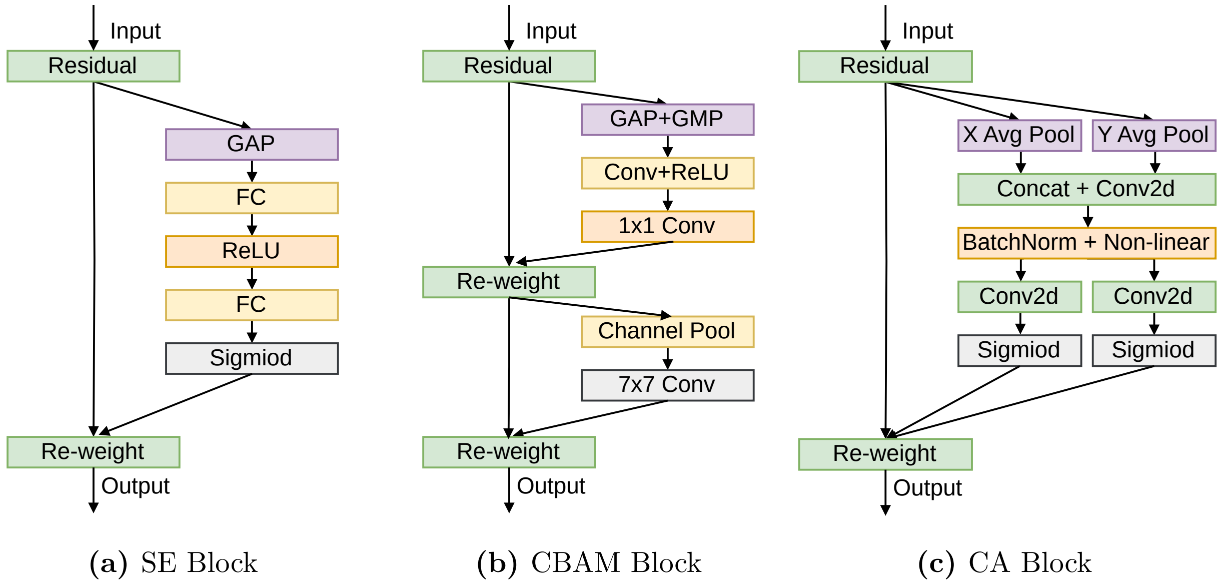
Figure 5.2: Visualization of blocks of (a) MobileNetV1 [74], and (b) MobileNetV2 [161].

decoupling the spatial filtering and feature combination steps seen in standard convolutions. This approach allows each input channel to be processed independently with its own filter, thereby focusing on extracting spatial features without mixing information across channels. The subsequent step in depth-wise separable convolutions involves combining these independently filtered channel outputs to construct new features achieved using pointwise convolutions.

5.4.1.2 MobileNet Architecture

Efficient network architecture refers to the design of neural network models optimized for performance and efficiency on mobile devices. These architectures are proposed to balance the trade-off between computational resources and performance which allow for real-time applications on hardware with limited processing power, memory, and energy consumption. The main building blocks of these architectures are depth-wise separable convolutions [73, 74, 161, 240]. Among the notable advancements in these architectures is the MobileNet which include MobileNetV1 and MobileNetV2. Each version represents an evolution in design to improve efficiency and accuracy as shown in Figure 5.2. MobileNetV1 is designed around the concept of depth-wise separable convolutions. This architecture significantly reduces the number of parameters and computational cost compared to standard convolutions, without a substantial decrease in model performance. Depth-wise separable convolutions divide the convolution operation into two layers: a depth-wise convolution that applies a single filter per input channel, and a pointwise convolution, a 1x1 convolution, that combines the output of the depth-wise convolution. This design makes MobileNetV1 highly efficient and suitable for mobile and embedded vision applications.

Building on the efficiency of MobileNetV1, MobileNetV2 introduces two key features include inverted residuals and linear bottlenecks. The architecture uses lightweight depth-wise convolutions to filter features in the intermediate expansion layer and em-



(a) SE Block **(b)** CBAM Block **(c)** CA Block

Figure 5.3: Various attention mechanisms used in the proposed MGAZE-Net include SE, CBAM and CA modules.

ploys pointwise convolution to project high-dimensional features into a low-dimensional space, which forms the bottleneck. This approach helps in reducing the representational bottleneck and improves the flow of information. Furthermore, the inverted residuals structure allows for the use of residual connections between the bottleneck layers, enhancing model training and performance. MobileNetV2 strikes an improved balance between latency and accuracy, making it more efficient than its predecessor. In MGAZE-Net, inverted residuals and linear bottlenecks from MobileNetV2 are utilized to tackle the computational cost problem of gaze estimation.

5.4.1.3 Attention Mechanisms

Attention mechanisms in computer vision are techniques that enable CNNs to selectively focus on specific parts of an input image, much like how humans focus on certain aspects of their visual environment. It allows the model to dynamically focus on the most informative parts of an input rather than treating all parts of the image equally. Various attention mechanisms are used in different computer vision tasks including SE module [77], CBAM [209] and CA [72] modules. These mechanisms are utilized effectively in various models, particularly in mobile networks, where they contribute significantly to improving state-of-the-art performance with a reduction in computational cost.

SE module enhances channel-wise feature responses by explicitly modeling interdependence between channels. It emphasizes the importance of different channels in a feature map by adaptively recalibrating channel-wise feature responses. As shown in Figure 5.3, SE block operates in two stages: squeeze and excitation modules. The squeeze module aggregates the global spatial information from the input feature map using global average pooling (GAP). This operation compresses each two-dimensional feature channel into a single scalar value. Following the squeeze operation, the excitation module captures channel-wise dependencies. It employs fully connected layers followed by ReLU and sigmoid activation to produce a channel-specific attention vector. This vector signifies the importance of each channel, enabling selective emphasis on more informative features.

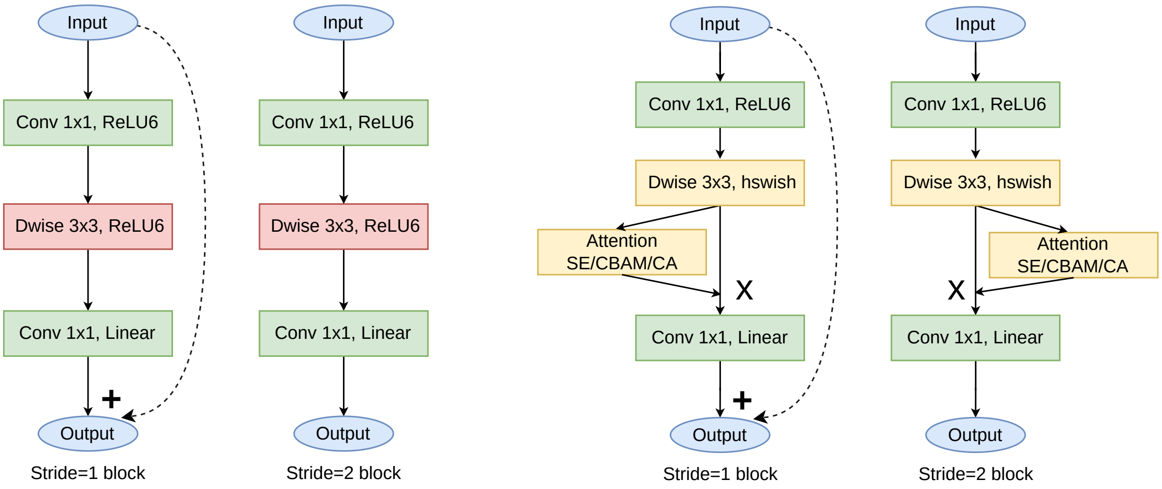
CBAM module enhances the representational power of CNNs by sequentially focusing on informative features across channel and spatial dimensions. This mechanism consists of two main components: channel attention and spatial attention, applied one after the other to refine the feature maps generated by convolutional layers. The channel attention module aims to emphasize meaningful features on a per-channel basis. It uses GAP and max pooling to generate two different spatial context descriptors, which capture the distinctive statistics of each channel. The spatial attention module focuses on where important features are located across the spatial dimension of the feature map. It employs both average-pooling and max-pooling across the channel axis to highlight salient spatial locations.

CA module introduces a coordinate-aware mechanism to capture long-range dependencies across spatial dimensions. It splits the channel attention into two 1D feature encoding processes for height and width, allowing the network to preserve important positional information alongside channel-wise dependencies. CA preserves crucial positional information by separately addressing the horizontal and vertical dimensions of the input feature map. It begins by dividing the input into two branches, each focusing on a different spatial dimension through dimension-specific global pooling. This results in two separate sets of features, each summarizing different spatial contexts. These features are then transformed through a shared network, typically involving convolutional and activation layers, to generate two distinct attention maps. Each map targets relevant features along its respective axis. The attention maps are subsequently expanded back to the dimensions of the original input and applied multiplicative, modulating the input based on the learned spatial attentions. This dual-axis focus allows CA to capture long-range dependencies more effectively, providing finer control over spatially relevant features and resulting in more precise and context-aware neural network performance.

In MGAZE-Net, attention mechanisms are used to improve the extraction of rich and fine-grained gaze features from face images.

5.4.2 MGAZE-Net Architecture

The architecture of MGAZE-Net focuses on accurately extracting gaze-related features from the limited eye region within large facial images. Additionally, it ensures computational efficiency suitable for real-time applications on mobile devices. This section explores the bottleneck and the overall network architecture of MGAZE-Net.



(a) MobileNetV2 bottleneck **(b)** MGN-block

Figure 5.4: Comparison between the bottleneck used in (a) MobileNetV2 and (b)MGAZE-Net. MobileNetV2 [161] bottlenecks, replacing ReLU with PReLU. MGAZE-Net utilizes MobileNetV2 bottleneck along with attention mechanisms including the SE, CBAM and CA. Attention modules are applied in the residual layer and followed by hswish as the nonlinearity.

5.4.2.1 MGAZE-Net Bottleneck

Attention mechanisms prioritize more informative parts of the input image, focusing less on the other regions in both the channel and the spatial dimensions. Further, MobileNet architecture is used to balance the trade-off between computational resources and performance, allowing real-time processing. The primary objective of MGAZE-Net is to assign significant weight to the eye regions, which hold the most crucial gaze information, while using efficient computational cost. To achieve this goal, MGAZE-Net bottleneck (MGN-block) is proposed to optimize the balance between performance and cost by integrating MobileNetV2 bottleneck and attention mechanisms.

MGN-block is built on the MobileNetV2 bottleneck as shown in Figure 5.4, which is known for its lightweight and efficient architecture. MobileNetV2 bottleneck utilizes inverted residuals and linear bottlenecks, which are particularly suited to managing the model's computational cost while preserving the richness of feature maps that are important for accurate predictions. For MGN-block, 1×1 pointwise convolution expands the input from a low-dimensional space to a higher-dimensional one. This is followed by a depth-wise convolution 3×3 that filters this expanded input. Finally, the features are compressed back into a low-dimensional space using another linear 1×1 pointwise convolution. This sequence efficiently manages feature dimensionality while preserving crucial information.

An attention layer is integrated directly after the depth-wise convolution in MGAZE-Net to emphasize crucial regions within the image. Three effective attention mechanisms are incorporated, including SE, CBAM, and CA, each selected for its unique

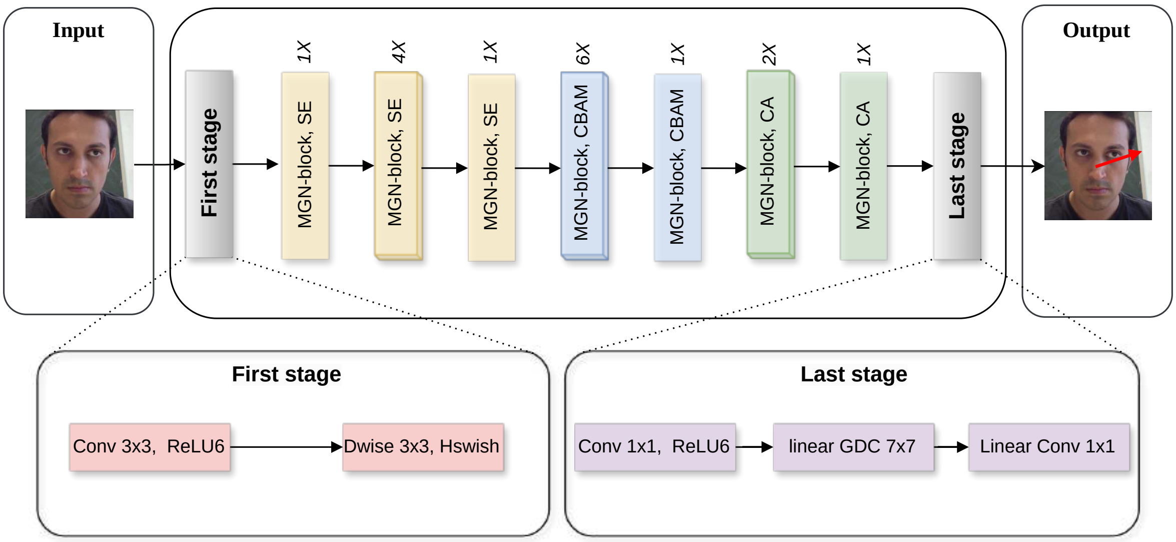


Figure 5.5: Architecture design of the proposed MGAZE-Net. The MGAZE-Net architecture incorporates 7 MGN-blocks with SE,CBAM and CA modules. This information is extracted from the middle blocks of the network.

strengths to improve the network's focus on key features related to gaze. Additionally, the h-swish non-linearity is introduced into the MGN-block. Compared to the standard ReLU non-linearity, h-swish has shown to significantly enhance neural network performance. However, this improvement incurs increased computational demands, particularly on embedded devices. To address this, h-swish is applied exclusively within the depth-wise convolutions where computational costs are lower. In contrast, ReLU6 is used in pointwise convolutions to preserve efficiency across MGAZE-Net. The integration of MobileNetV2's bottleneck with advanced attention mechanisms and h-swish non-linearity allows MGAZE-Net to effectively extract fine-grained gaze features while maintaining low computational demand.

5.4.2.2 Network Architecture

The entire MGAZE-Net architecture pipeline is presented in Figure 5.5. The first stage of MGAZE-Net starts with a fast downsampling stage to efficiently process input facial images. This stage consists of a standard 3 × 3 convolution layer followed by a 3 × 3 depth-wise convolution. This stage helps to reduce the spatial dimensions of the input while preserve essential features necessary for accurate gaze estimation. This is crucial for reducing spatial dimensions early and accelerating subsequent processing without significant loss of information.

Subsequently, the architecture incorporates seven layers of the proposed MGN-blocks. These bottlenecks use different strides (1 or 2) and integrated sequentially with different attention mechanisms include SE, CBAM, and CA to optimize the extraction of the most informative gaze features. The first three bottlenecks incorporate the SE attention mechanism, which improves the network's ability to capture important channel-wise relationships. The SE module helps the network focus on the most informative features by dynamically adjusting the weights of different channels according to their relevance. This selective emphasis on certain channels over others at the initial stages of the network ensures that subsequent layers process more refined and significant feature representations, which can lead to more accurate gaze estimation.

Following the SE, two bottlenecks employ CBAM attention mechanism, which utilize both channel and spatial attention. CBAM improves the quality of the feature representations that the network processes by analyzing both the inter-channel relationships and spatial dependencies within the feature maps. This is particularly beneficial after the network has already applied several layers of convolution and needs to refine these features for better accuracy. This feature refinement is important for handling variations due to occlusions, head poses, and different lighting conditions, thus increasing the robustness of the model in different environmental settings. The final two bottlenecks are equipped with CA attention mechanism, which capture long-range dependencies across the input feature maps. By the time the features have passed through initial layers and attention mechanisms like SE and CBAM, they have been sufficiently condensed and abstracted. Placing CA at the last bottlenecks allows the network to effectively utilize the rich, abstracted feature representations, integrating long-range spatial information, which is crucial for accurate gaze estimation.

By sequentially employing SE, CBAM, and CA in sequence within the MGN-blocks, the network benefits from a compounded refinement of attention. This progressive attention mechanism facilitates a more fine-grained feature extraction process, leading to substantial improvements in the accuracy of gaze estimation. GDConv is adapted in the last stage of MGAZE-Net. One of the most critical features of GDConv is its ability to assign different weights to specific units within the feature map. Unlike GAP, which uniformly processes all features equally, GDConv varies the influence of different parts of the feature map based on their information content. This selective weighting is crucial for gaze estimation, especially giving more focus to the eye regions where gaze information is critical. By focusing on these key areas, GDConv ensures that MGAZE-Net maintains its sensitivity to the most discriminative gaze features essential for determining accurate gaze direction.

5.4.2.3 Network Specifications

The network specifications of MGAZE-Net are shown in Table 5.1. The network begins with a standard convolutional layer that has a kernel 3×3 that applied to an input image of size 224×224. This layer expands the output channels to 64 and reduces the spatial dimensions to 112×112 using a stride of 2. It performs the initial feature extraction without any attention mechanism. This is followed by a 3×3 depthwise convolution that maintains the spatial dimensions and channel count, enhancing

Table 5.1: MGAZE-Net Specification: n refers to the number of repetitions, c refers to output channels, c refers to the channels, and s to the stride

Input	n	s	c	bottleneck	Attention
224 × 224	1	2	64	Conv 3 × 3	-
112 × 112	1	1	64	Depth-wise conv 3 × 3	-
112 × 112	1	2	64	MGN-block	SE
56 × 56	4	1	64	MGN-block	SE
56 × 56	1	2	128	MGN-block	SE
28 × 28	6	1	128	MGN-block	CBAM
28 × 28	1	2	128	MGN-block	CBAM
14 × 14	2	1	128	MGN-block	CA
14 × 14	1	2	256	MGN-block	CA
7 × 7	1	1	1024	Conv 1 × 1	-
7 × 7	1	1	-	Linear GDConv 7 × 7	-
1 × 6	1	1	-	Linear Conv 1 × 1	-

the network's ability to capture spatial hierarchies without increasing computational complexity. The first MGN-block incorporates SE module down samples the features to 56 × 56 with a stride of 2. This layer is repeated four times, each with a stride of 1. All iterations utilize the SE module while maintaining the channel size. This layers serve to refine channel-wise feature responses and emphasize important features. A final MGN-block with SE module is applied with a stride of 2 extending the output channels to 128 and reducing the spatial dimension to 28 × 28.

The attention mechanism switches to CBAM within the MGN-blocks. This shift occurs over several layers, progressively integrating spatial and channel-wise attention to enable more context-aware feature adjustments. At a reduced spatial dimension of 14 × 14, the network incorporates MGN-block with CA module to capture long-range dependencies across spatial dimensions which is crucial for accurately estimating gaze direction. The network ends with a convolution 1 × 1 to consolidate features in 7 × 7, followed by a GDConv using a kernel 7 × 7 to further reduce the spatial dimensions to 1 × 1. This unique final stage processes the entire spatial extent of the feature map, emphasizing regions critical to gaze estimation. Finally, a linear convolution with a 1 × 1 kernel outputs the predicted 6d gaze rotation matrix to effectively synthesize all refined features processed through the network into a final gaze estimation.

5.4.2.4 Gaze Representation and Loss Function

In MGAZE-Net, gaze representation is notably advanced by adopting a rotation matrix formalism as shown in Figure 5.6. This representation is designed to resolve the discontinuities and ambiguities commonly associated with spherical angle representa-

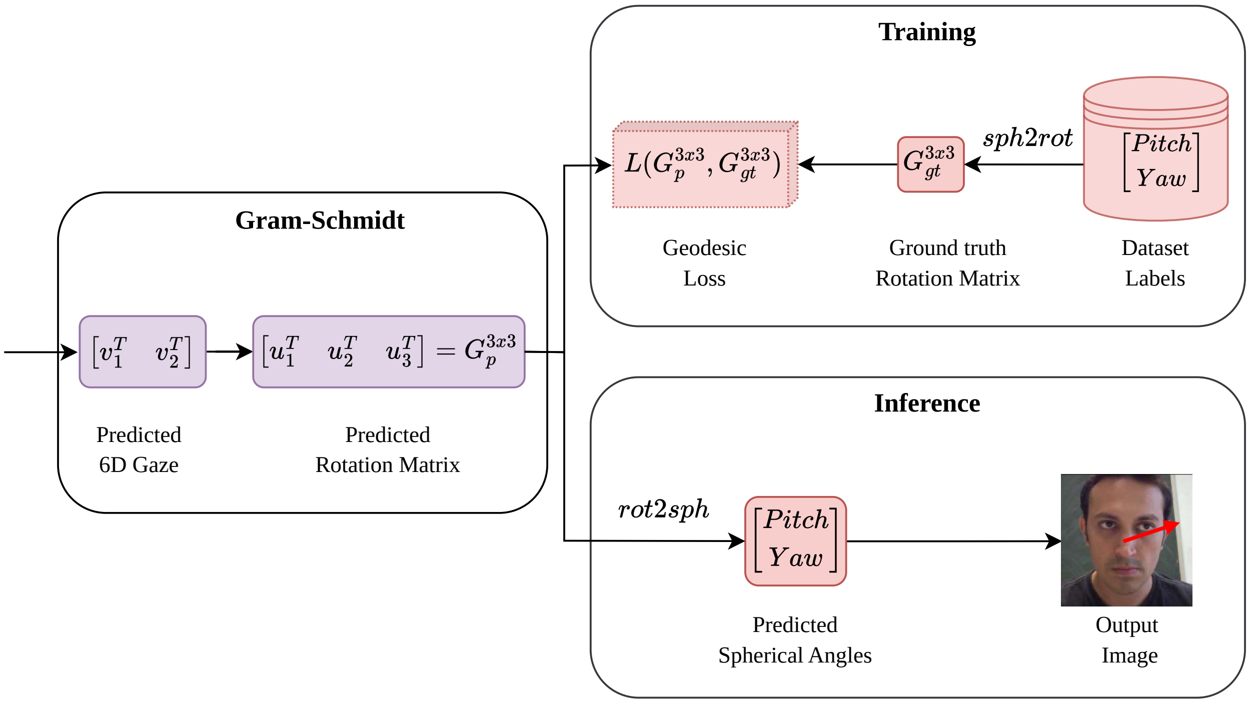


Figure 5.6: MGAZE-Net utilizes a gaze representation based on the rotation matrix using Gram-Schmidt and a geodesic-based loss function.

tions. MGAZE-Net further use the 6D rotation representation for efficient and direct regression. To effectively train MGAZE-Net with the rotation matrix representation, a geodesic loss function is employed. This loss function is formulated to compute the shortest path between the predicted and actual rotation matrices on the manifold of rotations. The geodesic loss between the predicted rotation matrix G_p and the ground truth G_{gt} is calculated using the formula:

$$\mathcal{L}_g(\theta_f, \theta_g) = \arccos\left(\frac{\mathrm{tr}(G_p G_{gt}^T) - 1}{2}\right). \tag{5.5}$$

Here, tr denotes the trace of a matrix, which is the sum of the diagonal elements, and $\cos^{-1}$ is the arc cosine function, providing the angle in radians that quantifies the rotational difference.

5.5 Experiments and Results

This section details the rigorous evaluation of MGAZE-Net that conducted to verify its performance and compare it with state-of-the-art gaze estimation methods. In this section, the experimental protocols (Section 5.5.1), quantitative (Section 5.5.2) and qualitative results (Section 5.5.3) are discussed.

5.5.1 Experimental Protocols

This section includes the experimental protocols required to assess the proposed framework, including evaluation datasets, performance metrics, data preprocessing, implementation details, and experiments setup.

5.5.1.1 Evaluation Datasets

Four large-scale datasets are used that reflect both controlled and unconstrained environments to train and test MGAZE-Net:

- **MPIIFaceGaze** [248]: offers 213,659 images from 15 subjects, recorded over several weeks of their daily activities. The wide variety of backgrounds and lighting conditions in this dataset makes it particularly suited for unconstrained gaze estimation tasks.
- **GazeCapture** [107]: is the largest available in-the-wild gaze dataset which contains a training set of 1,379,083 images, a testing set of 191,842 images, and a validating set of 63,518 images.
- **RT-GENE** [54]: offers 122,531 samples from subjects equipped with eye-tracking glasses. It provides high variability in distance from the camera, ranging from 0.5 to 2.9 meters.
- **Gaze360** [95]: This dataset features 172,000 images from 238 subjects over a wide spectrum of gaze and head poses, enhancing the robustness of our tests. For consistency, we focus on 84,900 images showing frontal faces.

5.5.1.2 Performance Metrics

The primary metric for assessing the effectiveness of gaze estimation models is the angular error between the estimated gaze direction and the ground truth in all test images. This error quantifies the accuracy of the model, with a smaller angle error indicating a more precise estimate. Given the ground truth gaze direction g and the predicted gaze direction ĝ, the angular error $\mathcal{L}_{angular}$ can be calculated as:

$$\mathcal{L}_{angular} = \arccos\left(\frac{\mathrm{g} \cdot \hat{\mathrm{g}}}{\|\mathrm{g}\|\|\hat{\mathrm{g}}\|}\right) \tag{5.6}$$

In addition to accuracy, the computational complexity of gaze estimation models is a critical metric, especially in real-time application scenarios. To objectively measure this aspect, extensive experiments across different methods are performed using the same hardware setup. This approach ensures a fair comparison of computational efficiency, highlighting the balance each model strikes between accuracy and speed.

5.5.1.3 Data Preprocessing

To prepare the images of the dataset for the gaze estimation model, normalization procedures are executed as outlined in previous research [245]. This process involves adjusting the virtual camera's position through rotation and translation to negate the roll angle of the head and maintain a consistent distance between the camera and the center of the face. This step is crucial for aligning the datasets with the input requirements of MGAZE-Net. For Gaze360, RT-Gene, and MPIIFaceGaze, the same procedures as in [34, 245] are followed to preprocess the dataset and create normalized face crops. For GazeCapture, the settings described in [231, 245] are used to create normalized head crops.

5.5.1.4 Implementation Details

MGAZE-Net is implemented in PyTorch and trained using the Adam optimizer at a learning rate of 0.0045. The training process uses 50 epochs with a batch size of 32. Notably, the model is trained from scratch on all datasets without relying on pretrained weights, ensuring that the results reflect the network's capabilities without external influences. The experiments are conducted using the PyTorch framework (version 1.8.1) on a computing setup equipped with an Intel(R) Core (TM) i7-7800X CPU, an NVIDIA RTX 3080, and 12GB of RAM.

5.5.1.5 Experiments Setup

To evaluate the robustness of MGAZE-Net, two training models are proposed. The first one named MGAZE-Net$V1$ is used to assess the accuracy of MGAZE-Net in within dataset evaluation. The second one named MGAZE-Net$V2$ is utilized to evaluate the generalization of MGAZE-Net in cross dataset evaluation.

MGAZE-NetV1 To evaluate the accuracy of MGAZE-Net, a detailed experiment is conducted that focuses on the evaluation of within-dataset in the four datasets mentioned. Each dataset involved different evaluation protocols based on previous research to ensure that the performance measures are reliable and comparable. Consequently, four within dataset evaluation tasks are established. In all tasks, MGAZE-Net is trained and tested on the same dataset with specified evaluation criteria include:

- $\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}$: In this task, 15-fold cross-validation as mentioned is used to evaluate MGAZE-Net on the MPIIFaceGaze dataset.
- $\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}$: In this task, a predefined split of 1379083 image training set, 191842 image testing set, and 63518 image validation set as specified by state-of-the-art.

- $\boldsymbol{\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}}$: In this task, a predefined split of 84000 training, 16500 testing, and 500 validation data as specified by the dataset authors are used.
- $\boldsymbol{\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}}$: In this task, 3-fold cross-validation is used to evaluate MGAZE-Net on RT-GENE dataset.

MGAZE-NetV2 To assess the reliability of MGAZE-Net, two experiments focusing on cross-dataset evaluations across the four mentioned datasets are conducted. The first experiment includes training MGAZE-Net simultaneously on Gaze360 ($\mathcal{D}_{GZ}$) and 300W-LP dataset ($\mathcal{D}_{WL}$) datasets. The second experiments include training MGAZE-Net simultaneously on RT-GENE ($\mathcal{D}_{RT}$) and 300W-LP dataset ($\mathcal{D}_{WL}$) datasets. Consequently, four cross-dataset evaluation tasks are established include:

- $\boldsymbol{\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}}$: In this task, MGAZE-Net is trained simultaneously on Gaze360 ($\mathcal{D}_{GZ}$) and 300W-LP dataset ($\mathcal{D}_{WL}$) and tested on MPIIFaceGaze ($\mathcal{D}_{MP}$).
- $\boldsymbol{\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}}$: In this task, MGAZE-Net is trained simultaneously on Gaze360 ($\mathcal{D}_{GZ}$) and 300W-LP dataset ($\mathcal{D}_{WL}$) and tested on GazeCapture ($\mathcal{D}_{GC}$).
- $\boldsymbol{\mathcal{D}_{RT} \rightarrow \mathcal{D}_{MP}}$: In this task, MGAZE-Net is trained simultaneously on RT-GENE ($\mathcal{D}_{RT}$) and 300W-LP dataset ($\mathcal{D}_{WL}$) and tested on MPIIFaceGaze ($\mathcal{D}_{MP}$).
- $\boldsymbol{\mathcal{D}_{RT} \rightarrow \mathcal{D}_{GC}}$: In this task, MGAZE-Net is trained simultaneously on RT-GENE ($\mathcal{D}_{RT}$) and 300W-LP dataset ($\mathcal{D}_{WL}$) and tested on GazeCapture ($\mathcal{D}_{GC}$).

5.5.2 Quantitative Results

This section presents the quantitative results of the experiments, demonstrating the superiority of MGAZE-Net in improving accuracy, reliability and efficiency over existing state-of-the-art techniques.

5.5.2.1 Within Dataset Evaluation

MGAZE-NetV1 is evaluated on the four within dataset tasks including $\boldsymbol{\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}}$, $\boldsymbol{\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}}$, $\boldsymbol{\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}}$ and $\boldsymbol{\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}}$. As mentioned in Section 5.5.1.5, each task is evaluated based on specific criteria to provide a fair comparison with state-of-the-art methods. MGAZE-NetV1 is robustly evaluated against a wide range of state-of-the-art gaze estimation methods. The comparison includes various methods and network architectures, highlighting the diversity and evolution in the field. Further, it present traditional methods that utilize both face and eye images to estimate gaze direction such as DilatedNet, FullFace, CA-Net, RT-Gene, AGE-Net, Fare-Net, and L2DNet. These methods require a high computational cost, as they utilize multiple backbones to extract eye and face features. Additionally, it examines methods

Table 5.2: Comparison with the state-of-the-art methods in within dataset evaluation: MGAZE-NetV1 achieves state-of-the-art gaze performance in the four datasets evaluation tasks with mean angular errors of 3.88°, 2.87°, 10.48°, and 6.52°on $\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}$, $\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}$, $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}$ and $\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}$, respectively.

Methods	$\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}$ (15-fold)	$\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}$ (test-set)	$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}$ (test-set)	$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}$ (3-fold)
FullFace [251]	4.90°	-	-	7.44°
AGE-Net [14]	4.09°	-	-	7.44°
GazeTR-Pure [32]	4.74°	-	13.58°	8.06°
L2DNet [133]	4.30°	-	-	8.40°
Fare-Net [36]	4.30°	-	-	8.40°
SAtten-Net [135]	4.04°	-	10.70°	7.00°
CA-Net [30]	4.27°	-	11.20°	8.27°
RT-Gene [54]	4.30°	-	12.26°	8.00°
MANNet [215]	4.30°	-	-	13.20°
ETH-Gaze [242]	4.80°	3.30°	-	12.00°
Gaze360 [95]	4.06°	-	11.20°	7.12°
FAZE [144]	-	3.49°	-	-
RSN [247]	4.50°	3.32°	-	-
GEDDNet [28]	4.69°	-	-	8.17°
Dilated-Net [27]	4.42°	-	13.73°	8.38°
EM-Gaze [215]	4.10°	-	-	-
MGAZE-NetV1	**3.88°**	**2.87°**	**10.48°**	**6.52°**

such as Gaze360, ETH-Gaze, and SAtten-Net, which rely exclusively on facial images. These methods reflect a shift towards more efficient models that maintain reasonable accuracy while simplifying input requirements. The methods also vary in architectural design, from standard CNNs and dilated CNNs to more sophisticated designs featuring transformers with self-attention mechanisms. This diverse comparison to effectively evaluate the efficiency and performance of MGAZE-NetV1 across challenging datasets.

The results in Tables 5.2 provide a comprehensive comparison between MGAZE-NetV1 and other state-of-the-art methods. For a fair comparison, the results of all methods are reported either directly from their original papers or derived by executing their publicly available codes. MGAZE-NetV1 achieves the lowest mean angular error in all tasks. Further, MGAZE-NetV1 outperforms the state of the art reported performance by approximately 4%, 13%, 2% and 3% in the tasks $\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}$, $\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}$, $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}$ and $\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}$, respectively. These findings confirm the superior capability of MGAZE-NetV1, illustrating its effectiveness in leveraging fine-

Table 5.3: Comparison with the state-of-the-art methods in cross dataset evaluation: MGAZE-NetV2 achieves state-of-the-art gaze performance in the four cross datasets tasks with mean angular errors of 6.15°, 15.55°, 8.45°, and 11.85° on $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}$, $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}$, $\mathcal{D}_{RT} \rightarrow \mathcal{D}_{MP}$ and $\mathcal{D}_{RT} \rightarrow \mathcal{D}_{GC}$, respectively.

Category	Methods	$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}$	$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}$	$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{MP}$	$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{GC}$
Typical Methods	Full-Face [251]	13.53°	22.23°	14.40°	18.16°
	RTGene [54]	14.52°	20.11°	10.95°	15.62 °
	CA-Net [30]	16.27°	22.11°	15.62°	18.23°
	Dilated-Net [27]	11.72°	18.9°	8.92°	14.21°
	Eth-Gaze [242]	10.3°	16.98°	12.0°	13.2°
	Gaze360 [95]	8.15°	17.80°	9.12°	13.58°
Domain Generalization Methods	PureGaze [29]	9.28°	17.22°	-	-
	SelfAtt [135]	8.31°	17.70°	-	-
	MGAZE-NetV2	**6.15°**	**15.55°**	**8.45°**	**11.85°**

grained gaze features to surpass state-of-the-art models. The integration of lightweight CNN architectures with attention mechanisms allows MGAZE-Net to maintain low computational demands while enhancing the accuracy of gaze estimation.

5.5.2.2 Cross-Dataset Evaluation

MGAZE-NetV2 is extensively evaluated on the four cross dataset evaluation tasks include $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}$, $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}$, $\mathcal{D}_{RT} \rightarrow \mathcal{D}_{MP}$ and $\mathcal{D}_{RT} \rightarrow \mathcal{D}_{GC}$. As mentioned in Section5.5.1.5, each task is trained on a specific dataset and tested against another to provide a robust evaluation of the performance of gaze generalization. MGAZE-NetV2 is rigorously evaluated against a wide range of state-of-the-art gaze estimation methods. The comparison includes various methods and network architectures, highlighting the diversity and evolution of the field. Furthermore, it presents typical gaze estimation methods such as Full-Face [251], RT-Gene [54], CA-Net [30], Dilated-Net [27], Eth-Gaze [242], and Gaze360 [95]. These methods generally aim to enhance within dataset performance through comprehensive network designs and approaches. However, they often fail in cross-dataset scenarios due to the overfitting to specific domain features. Additionally, it examines domain generalization methods like SelfAtt [135] and PureGaze [29] which designed to enhance cross-dataset performance without incorporating target domain samples. These approaches generally exhibit improved performance over typical gaze estimation methods, but still struggle to completely overcome the limitations imposed by source domain bias.

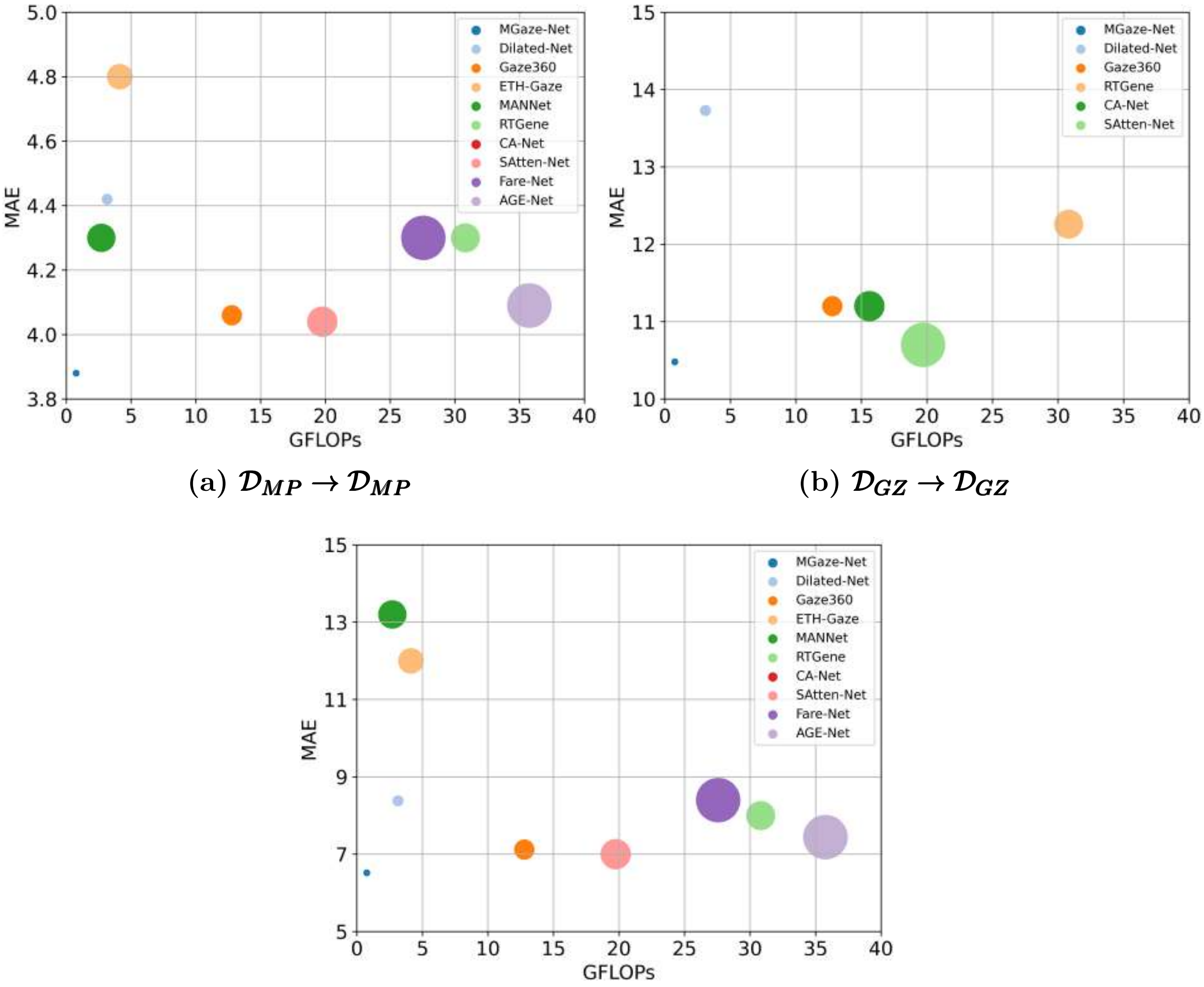
The comparative analysis in Table 5.3 reveals that the MGAZE-NetV2 framework outperforms typical gaze estimation methods and domain generalization methods. The results demonstrate that MGAZE-NetV2 is capable of generalizing accurately across different domains and significantly improving the reliability of gaze estima-

Table 5.4: Comparison with the state-of-the-art methods in terms of computational complexity: MGAZE-Net achieves lowest parameter count, FLOPs, and time compared with the top-performing compact models from recent literature on each task.

Methods	**Backbone**	**#Params** (M)	**#FLOPs** (G)	**Time** (ms)
FullFace [251]	CNN	196.6	2.99	-
AGE-Net [14]	Tr	109.0	35.75	66
GazeTR-Pure [32]	Tr	104.0	58.3	-
L2DNet [133]	Dilated-CNN	87.0	-	-
Fare-Net [36]	CNN	75.0	27.5	49
SAtten-Net [135]	Tr	74.8	19.7	379
CA-Net [30]	CNN	34.0	15.6	30
RT-Gene [54]	CNN	31.0	30.81	35
MANNet [215]	CNN	29.5	2.7	-
ETH-Gaze [242]	CNN	23.8	4.12	6.4
Gaze360 [95]	RNN	14.6	12.78	5.4
CDBN [270]	CNN	13.0	-	8.4
FAZE [144]	CNN	-	-	-
RSN [247]	CNN	-	-	-
GEDDNet [28]	Dilated-CNN	4.0	-	6.3
Dilated-Net [27]	Dilated-CNN	3.9	3.1	6.7
EM-Gaze [215]	CNN	2.7	-	7.3
MGAZE-Net	**CNN**	**1.3**	**0.75**	**3.4**

tion in unseen datasets. As detailed in Table 5.3, typical gaze estimation methods often perform poorly in cross-dataset evaluations, since they are limited by source domain distribution and quality. Furthermore, domain generalization approaches report better gaze performance compared to typical gaze estimation methods. However, these improvements are still not sufficient, as these methods are prone to overfitting in the source domains. In contrast, MGAZE-NetV2 framework demonstrates superior cross-dataset gaze performance and achieves the best results across all tasks. MGAZE-NetV2 outperforms the last reported state-of-the-art by approximately 25%, 8%, 4% and 10% on $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{MP}$, $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GC}$, $\mathcal{D}_{RT} \rightarrow \mathcal{D}_{MP}$ and $\mathcal{D}_{RT} \rightarrow \mathcal{D}_{GC}$, respectively. This achievement can be attributed to the strong representation of features and the improved knowledge sharing achieved through the attention mechanisms and the multitask approach.

Figure 5.7: Illustrating the Balance: Computational Complexity, Performance, and Model Size Across Benchmarks. Model size is proportionally represented by marker area. MGAZE-Net, highlighted with a blue circle, exemplifies a leading trade-off between gaze estimation performance, FLOPs, and compactness, consistently occupying the bottom-left corner and compared with the top-performing compact models from recent literature on each benchmark.



(a) $\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}$

(b) $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}$

(c) $\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}$

5.5.2.3 Computational Complexity

To assess the computational complexity of MGAZE-Net, various state-of-the-art gaze estimation methods are evaluated in terms of processing time, parameter count, and floating point operations per second (FLOPs). Table 5.4 shows the computational complexity analysis between MGAZE-Net and the state of the art methods. The methods are categorized based on their model complexities into three different groups: those with over 30 million parameters, those ranging between 5 to 30 million parameters, and compact models with fewer than 5 million parameters. The results, as summarized in Table 5.4 reveal a distinct advantage in computational efficiency for

CNN methods over their counterparts that utilize transformer models. For instance, transformer models such as AGE-Net [14], GazeTR-Pure [32] and SAtten-Net [135] require high computational complexity compared with CNNs. In addition, the table highlights the effectiveness of MGAZE-Net in computational efficiency. With only 1.3 million parameters and 0.75 GFLOPs, MGAZE-Net is significantly faster than other state-of-the-art methods, requiring only 3.4 milliseconds to process the input image. This contrasts with heavier models like FullFace and GazeTR-Pure, which have larger parameter counts and computational costs.

To comprehensively assess the efficiency and accuracy of gaze estimation models, a scatter plot (referenced as Figure 5.7) is used. This plot maps the computational complexity against gaze accuracy for various state-of-the-art methods across different datasets. In this graph, each method is represented as a point (x, y) on the graph: x corresponds to GFlops, y corresponds to the mean angular error and the size of the point is proportional to the model's parameter count multiplied by 10. This scaling visually clarifies that models with larger parameter counts, reflecting their potential complexity and memory requirements. The models in the lower left corner of the graph, which include lower GFlops and lower mean angular errors, represent an optimal balance between computational efficiency and high accuracy. Remarkably, MGAZE-Net consistently occupies the lower left corner compared to the other methods, demonstrating a state-of-the-art trade-off between gaze performance and model compactness. This indicates that MGAZE-Net achieves an optimal trade-off between model complexity and computational efficiency.

The results in Table 5.2, Table 5.3, and Table 5.4 confirm the superior capability of MGAZE-Net, illustrating its effectiveness in leveraging fine-grained gaze features to surpass state-of-the-art models while providing the lowest computational cost against top methods. These findings highlight the effectiveness of the proposed method in balancing model complexity with computational efficiency, making it a good solution for real-time applications on mobile and embedded systems where computational resources are limited.

5.5.2.4 Ablation Study

a comprehensive ablation study is conducted to validate the individual contributions of different components within MGAZE-Net framework, specifically focusing on network architecture, and gaze representation. The evaluation is performed through systematic experiments on $\boldsymbol{\mathcal{D}_{MP}} \rightarrow \boldsymbol{\mathcal{D}_{MP}}$, $\boldsymbol{\mathcal{D}_{GC}} \rightarrow \boldsymbol{\mathcal{D}_{GC}}$, $\boldsymbol{\mathcal{D}_{GZ}} \rightarrow \boldsymbol{\mathcal{D}_{GZ}}$ and $\boldsymbol{\mathcal{D}_{RT}} \rightarrow \boldsymbol{\mathcal{D}_{RT}}$ tasks.

1) Network architecture: First, an ablation study is conducted to assess the individual and combined contributions of the attention mechanisms to the performance of MGAZE-Net. These mechanisms are evaluated through their integration into the network's bottleneck layers, as detailed in Table 5.5. The first row of the table repre-

Table 5.5: Effects of different attention mechanisms on the performance of MGAZE-Net framework. The last row indicates the MGAZE-Net performance results.

SE	CBAM	CA	$\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}$	$\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}$	$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}$	$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}$
✗	✗	✗	4.45°	4.43°	11.43°	7.20°
✗	✗	✓	4.56°	3.72°	11.02°	6.89°
✗	✓	✗	4.81°	3.90°	10.90°	6.95°
✓	✗	✗	4.25°	3.63°	10.63°	6.64°
✗	✓	✓	3.98°	2.95°	10.57°	6.69°
✓	✗	✓	4.08°	3.14°	10.74°	6.82°
✓	✓	✗	4.00°	3.02°	10.62°	6.65°
✓	✓	✓	**3.88°**	**2.87°**	**10.48°**	**6.52°**

Table 5.6: Ablation study on network architecture by systematically exchangingGDConv with GAP layer.

Methods	$\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}$	$\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}$	$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}$	$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}$
MGAZE-Net(w GAP)	4.20°	3.98°	10.72°	6.75°
MGAZE-Net	**3.88°**	**2.87°**	**10.48°**	**6.52°**

sents the scenario where no attention mechanisms are used, serving as a baseline. The next three rows detail the performance when only one of the mechanisms (SE, CBAM, or CA) is employed, with the others omitted. Subsequent rows examine the effects of combining two mechanisms, while excluding the third, in various configurations. The final row, representing the complete MGAZE-Net with all three mechanisms integrated. As summarized in Table 5.5, the final row shows the superior performance with the lowest angular error. This outcome validates that employing progressive attention mechanisms that begin with SE blocks, advance through CBAM, and end with CA effectively highlight the most important features and provide accurate gaze estimation.

Secondly, another experiment assesses the impact of the GDConv layer on the performance of MGAZE-Net. In this test, the GDConv is replaced with the traditional GAP layer typically used in CNN architectures. The findings presented in Table 5.6 indicate a significant enhancement in performance when using GDConv compared to GAP. This result confirms the effectiveness of GDConv in selectively emphasizing informative regions within the last feature map which is crucial for accurate gaze estimation.

2) Gaze rotation representation: In this study, an ablation analysis was conducted to evaluate the importance of the 6D gaze representation within MGAZE-Net. The

Table 5.7: Ablation study on gaze representation shows that using rotation representation improves the gaze performance on the four tasks compared with spherical angles.

Methods	$\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}$	$\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}$	$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}$	$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}$
Spherical angles	4.80°	3.86°	11.45°	6.75°
Rotation matrix	**3.88°**	**2.87°**	**10.48°**	**6.52°**

baseline MGAZE-Net served as the benchmark which employs the 6D gaze representation. To investigate the impact of this representation, a variant of the network was devised that omits the 6D representation and instead directly regresses the spherical angles. The experiments are performed on the same datasets, consistent hyperparameters, and identical training procedures as the baseline. The results emphatically demonstrate the superiority of the 6D representation, with the baseline MGAZE-Net outperforming the spherical coordinates model by a significant margin of up to 8%, as detailed in Table 5.7. This finding highlights the effectiveness of the 6D gaze representation in achieving precise and continuous gaze estimation, affirming its utility in the model and its potential applicability in other gaze estimation frameworks.

Overall, the ablation study underscores the critical role of the RCS-Loss function, the dual fully connected layer architecture, and the EfficientNet backbone in pushing the limits of gaze estimation performance. These components collectively contribute to our model's state-of-the-art accuracy, as evidenced by the substantial improvements across various datasets.

5.5.3 Qualitative Results

This section presents the qualitative results of using MGAZE-Net on real data. This provides invaluable insights into the model's practical capabilities and the nuances of its predictions.

5.5.3.1 Visualization of Gaze Predictions

A series of visualizations showing the gaze predictions of MGAZE-Net are presented in Figure 5.8. These visualizations highlight the robustness of MGAZE-Net and its ability to accurately estimate gaze direction despite potential challenging factors including changes in lighting, subject poses, and complex backgrounds. Furthermore, the visual alignment between the predicted gaze vectors and the actual gaze directions of the subjects emphasizes the model's effectiveness in capturing the intricacies of human gaze.

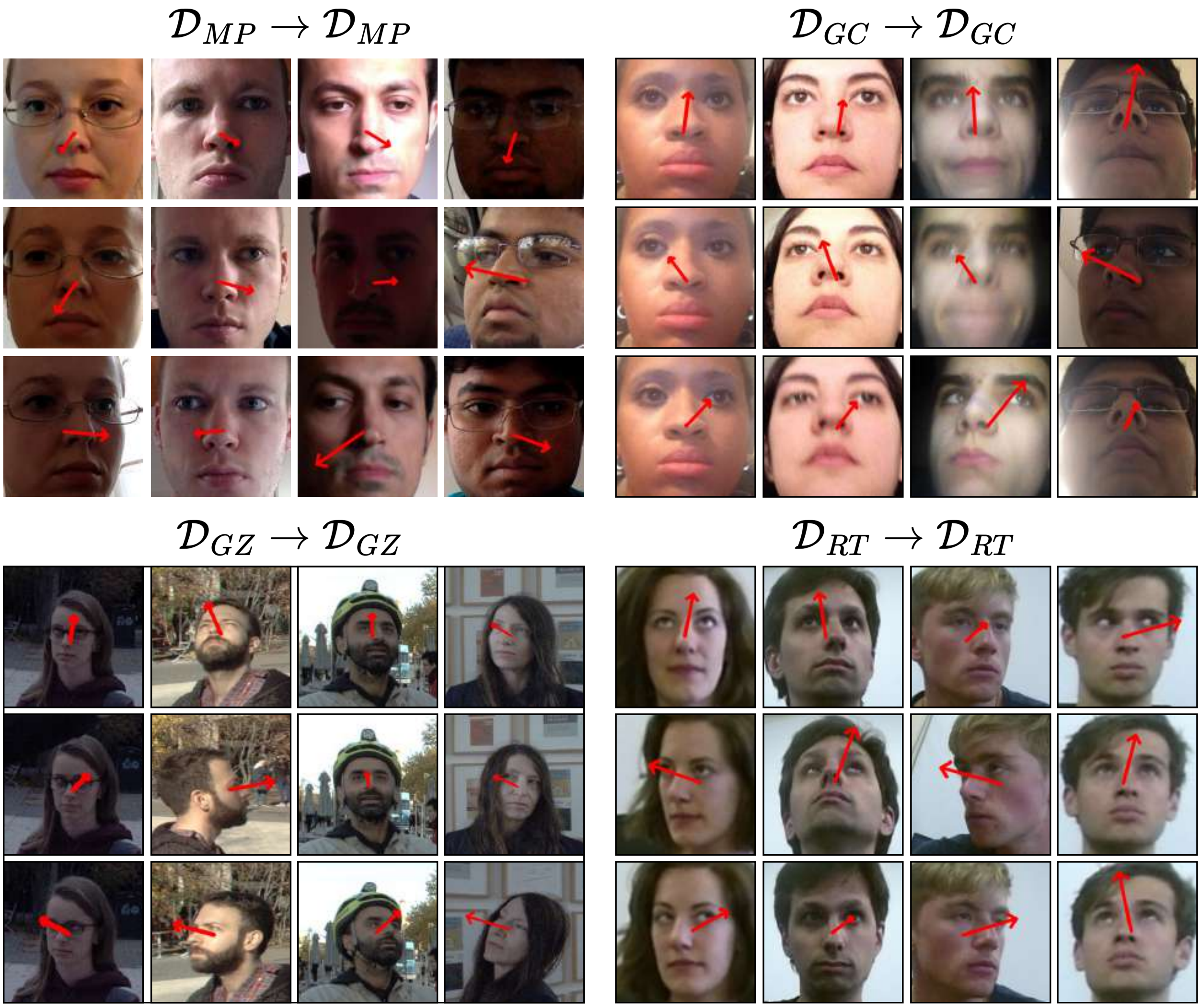

$\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}$
$\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}$
$\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}$
$\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}$

Figure 5.8: Visualization of Gaze Predictions of MGAZE-Net on face images on the four within dataset tasks include $\mathcal{D}_{MP} \rightarrow \mathcal{D}_{MP}$, $\mathcal{D}_{GC} \rightarrow \mathcal{D}_{GC}$, $\mathcal{D}_{GZ} \rightarrow \mathcal{D}_{GZ}$ and $\mathcal{D}_{RT} \rightarrow \mathcal{D}_{RT}$. These images are selected to show a variety of subjects, backgrounds, lighting, and camera angles.

5.5.3.2 Visualization of the Learned Features

The effectiveness of adapting attention mechanisms in the network is further analyzed using the class activation map [259] by visualizing the learned features of the model with and without attention mechanisms. These visualizations highlight the allocation of attention towards the input data, with regions of warmer color (e.g., red) indicating a greater degree of attention, while cooler regions reflect less attention. As shown in Figure 5.9, the visualizations clearly demonstrate that MGAZE-Net with attention mechanisms successfully assigns greater importance to the regions of the eye, especially in challenging conditions. In contrast, the model without attention mechanisms fails to capture crucial eye features under challenging lighting conditions. These results illustrate the effectiveness of the attention mechanisms employed in the model for

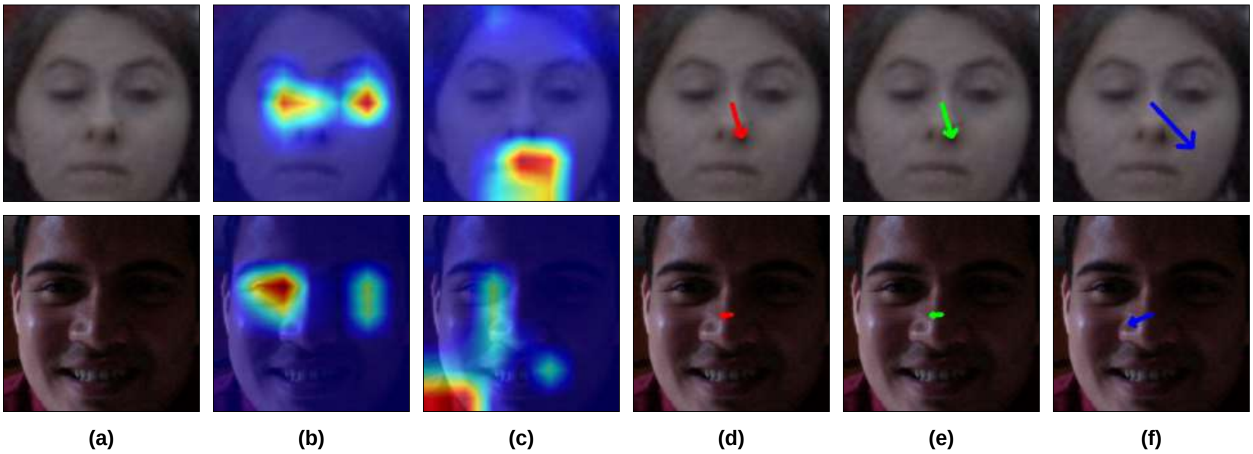


Figure 5.9: Grad-CAM [259] visualization of typical MGAZE-Net and MGAZE-Net without using attention mechanisms. (a) Input images, (b) Grad-CAM of typical MGAZE-Net, (c) Grad-CAM of MGAZE-Net without using attention mechanisms, (d) Ground truth, (e) Prediction from typical MGAZE-Net, (f) Prediction from MGAZE-Net without using attention mechanisms.

extracting gaze-related features from facial images. Additionally, gaze predictions of the model, both with and without attention mechanisms, are visualized alongside ground-truth values in the same Figure 5.9. This figure shows that MGAZE-Net is capable of generating accurate gaze predictions and performs better in gaze estimation.

5.6 Conclusion

This chapter introduced MGAZE-Net, a novel approach to improve the balance between gaze performance and computational resources tackle the challenges posed by capturing fine-grained gaze features in the constrained eye region of face images with low computational resources. MGAZE-Net integrates inverted residuals and linear bottlenecks from MobileNetV2 with advanced attention modules include SE, CBAM, and CA modules. This architecture enables MGAZE-Net to capture fine-grained gaze features in the constrained eye region of face images with low computational resources. Moreover, MGAZE-Net employs a rotation matrix formalism with a geodesic-based loss function for gaze representation. This approach mitigates issues related to discontinuity and ambiguity found in spherical angle representations. It provides a continuous and geometrically accurate method for representing gaze direction, improving learning efficiency and gaze estimation accuracy. Extensive experiments on four challenging datasets validated the efficiency of MGAZE-Net. The results demonstrated a state-of-the-art balance between performance and computational cost.

6 Conclusion and Future Work

This thesis has addressed the complex challenges of gaze estimation in unconstrained environments through the development and validation of innovative deep learning architectures and methodologies. The research carried out has significantly advanced the state of the art in gaze estimation technology, providing robust solutions that improve accuracy, reliability, and efficiency in real-world applications.

6.1 Summary of Contributions

This section provides a summary of the significant contributions made by this dissertation in the field of gaze estimation. Each contribution has been instrumental in advancing the understanding and application of gaze estimation technologies, especially in unconstrained environments.

6.1.1 A Survey of Deep Learning-Based Gaze Estimation

This thesis begins with a comprehensive review of the current deep learning approaches in gaze estimation, focusing on how these methodologies have evolved to tackle the complex problem of accurate gaze prediction in real-world settings. The survey highlights the shift from traditional model-based and feature-based methods to deep learning techniques, which are capable of handling high-dimensional data and offer significant improvements in accuracy and reliability under varied environmental conditions. It explores the application of various architectures, from CNNs to more advanced RNNs and attention mechanisms, demonstrating how each has contributed to overcoming the limitations of earlier methods. Furthermore, the review discusses the importance of robust and diverse datasets that enable these models to perform well in a range of scenarios, emphasizing the transition towards more adaptable and generalizable systems. This detailed analysis not only sets the foundation for the subsequent chapters, but also outlines the critical advances and ongoing challenges in the field of gaze estimation.

6.1.2 L2CS-Net, Fine-grained gaze estimation using multi-loss approach.

The primary objective of L2CS-Net is to address the challenge of extracting fine-grained gaze features from the small eye region within face images. To achieve this, a novel two-branch CNN architecture has been developed, specifically designed to predict pitch and yaw gaze angles separately. This model employs a multi-loss approach that integrates both classification and regression losses, significantly enhancing the accuracy of gaze direction estimation. By isolating the prediction of each gaze angle into dedicated fully connected layers, L2CS-Net is able to focus more effectively on learning discriminative features specific to each angle, thus improving the accuracy of gaze estimation. The multi-loss approach further strengthens the model by utilizing classification for coarse gaze direction estimation and regression for fine-grained predictions, achieving a higher level of accuracy.

The effectiveness of this model has been demonstrated through extensive evaluations across four popular gaze estimation datasets. MPIIFaceGaze, GazeCapture, Gaze360, and RT-GENE. This evaluation reveals that L2CS-Net achieves state-of-the-art performance, showing its exceptional capability to accurately estimate gaze direction across diverse environments. Comprehensive ablation studies have also been conducted to validate the effectiveness of the proposed network architecture and loss function. These studies highlight the importance of individual gaze angle prediction and the synergistic benefits of combining classification and regression losses, confirming the model's innovative approach to improving gaze estimation accuracy.

6.1.3 MTGH-Net: Multi-task gaze and head pose estimation

The primary objective of MTGH-Net is to improve the reliability of gaze estimation in different domains. To address this, an innovative framework leverages the intrinsic relationship between gaze and head pose estimation through a multi-task learning paradigm. By integrating gaze and head pose estimation into a single network, MTGH-Net achieves synergistic improvements in generalization capability for both tasks. MTGH-Net tackles the challenge of gaze generalization by employing a novel training strategy that utilizes two separate datasets one for gaze and another for head pose. This approach enriches the model's exposure to diverse conditions, significantly enhancing its ability to generalize across datasets with varying features. Additionally, MTGH-Net introduces a simplified yet effective 6D-parameter rotation matrix representation for both gaze and head pose estimation tasks. This, coupled with a geodesic-based loss function, facilitates an accurate and direct regression of these tasks. This approach effectively address the discontinuity issues inherent in traditional gaze representation methods and preventing bias towards either task.

The effectiveness of MTGH-Net has been validated through a detailed evaluation and benchmarking against current state-of-the-art methods. MTGH-Net exceeds the performance of single-task gaze networks, demonstrating up to a 21% improvement in gaze generalization performance on popular benchmarks. This comprehensive evaluation underscores the practical applicability and robustness of the framework in real-world scenarios. An extensive ablation study included in the MTGH-Net evaluation provides insightful analyses into the impact of its various components on overall performance. This study offers valuable insights into the contributions of the multitask learning approach, the continuous 6D representation, and the geodesic-based loss function. By analyzing the framework's components and assessing their individual and collective effects, the research enriches the understanding of effective strategies for gaze and head pose estimation.

6.1.4 MGAZE-Net: Robust gaze estimation mobile network

The primary objective of this research is to address the high computational costs associated with traditional gaze estimation models, which hinder their applicability in real-time scenarios. To overcome this, a novel and lightweight CNN architecture, MGAZE-Net, has been developed, augmented with a progressive combination of attention mechanisms, including SE, CBAM, and CA mechanisms. This hierarchical integration of attention mechanisms is designed to systematically emphasize crucial gaze information by capturing both local and global spatial relationships within facial images. The strategic placement of these mechanisms enables MGAZE-Net to efficiently extract fine-grained features critical for accurate gaze estimation, significantly reducing the computational overhead typically associated with deep CNN models and transformers.

MGAZE-Net employs a rotation matrix formalism to represent gaze direction, which mitigates the issues of discontinuity and ambiguity commonly found with spherical angle representation. This approach provides a continuous, unambiguous, and geometrically precise method for gaze representation. Additionally, MGAZE-Net introduces a geodesic loss function that leverages the geometric properties of rotation matrices. This loss function provides a more accurate measure of the discrepancy between predicted and actual gaze values, facilitating more effective training by directly penalizing errors in the rotation space, which substantially enhances the model's performance.

The efficacy and robustness of MGAZE-Net have been rigorously validated through extensive experiments on four challenging datasets: MPIIFaceGaze, GazeCapture, Gaze360, and RT-GENE. The comprehensive evaluation demonstrates that MGAZE-Net outperforms existing state-of-the-art methods across these datasets, highlighting its potential as a superior solution for real-time gaze estimation applications.

6.2 Future Directions

This thesis identifies several avenues for extending research in gaze estimation. The following directions are proposed to enhance the performance and applicability of gaze estimation technologies:

Large-Scale Gaze and Head Pose Dataset: Creating a comprehensive dataset that integrates gaze data with head pose annotations will allow for more accurate modeling and analysis of these interconnected behaviors. This integration will facilitate the development of models that can simultaneously predict both gaze and head pose with higher performance. Automatically annotating such datasets will streamline the process, ensuring that large scales of data can be processed efficiently. Furthermore, this dataset will serve as a valuable resource for the research community, promoting further innovation and testing of new algorithms. The aim is to capture a wide array of behaviors in various environmental settings to reflect real-world complexities. Ultimately, this endeavor will push the boundaries of what is currently possible in gaze and head pose estimation technologies.

Multi-Task Learning Approaches: The integration of multi-task learning models that can process additional features such as facial expressions and eye contact along with gaze and head pose holds significant potential. This approach can lead to the development of a unified model capable of extracting and utilizing correlations between these features to enhance overall prediction accuracy. By training models on multiple related tasks, the network can improve its generalization capabilities, as learning to predict one feature can help in predicting another. Multi-task learning is particularly effective in preventing overfitting, as it regularizes the model by sharing representations between related tasks. Additionally, such models are more efficient as they can perform multiple tasks simultaneously, reducing the need for separate models. Future research could explore various architectures and training strategies to optimize these models for speed and accuracy. Implementing such approaches could revolutionize user interaction technologies, making them more intuitive and responsive.

Learning with Less Supervision: The evolution of gaze estimation methods towards less supervised learning paradigms is crucial for their adaptation to diverse and uncontrolled environments. Exploring unsupervised, self-supervised, and weakly supervised learning methods will help in reducing the reliance on meticulously annotated data, which is costly and time-consuming to produce. These techniques, which leverage unlabeled data, could uncover new insights into gaze behavior by learning from a broader range of natural interactions without the constraint of label availability. Moreover, less supervised methods could enhance the robustness of gaze estimation systems, making them more flexible and scalable. Investigating these methods could also lead to better ways of handling the inherent noise and variability in real-world data, thereby improving the practicality and reliability of gaze estimation applications. The push towards less supervised learning models could ultimately make gaze estimation technology more accessible and applicable across different fields and applications.

Bibliography

[1] J. A. Abbasi, D. Mullins, N. Ringelstein, P. Reilhac, E. Jones, and M. Glavin. An analysis of driver gaze behaviour at roundabouts. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):8715–8724, 2021.

[2] A. A. Abdelrahman, D. Strazdas, A. Khalifa, J. Hintz, T. Hempel, and A. Al-Hamadi. Multi-modal engagement prediction in multi-person human-robot interaction. *IEEE Access*, 2022.

[3] A. A. Abdelrahman, D. Strazdas, A. Khalifa, J. Hintz, T. Hempel, and A. Al-Hamadi. Multimodal engagement prediction in multiperson human–robot interaction. *IEEE Access*, 10:61980–61991, 2022.

[4] N. Aghli and E. Ribeiro. A data-driven approach to improve 3d head-pose estimation. In *Advances in Visual Computing: 16th International Symposium, ISVC 2021, Virtual Event, October 4-6, 2021, Proceedings, Part I*, pages 546–558. Springer, 2021.

[5] A. A. Akinyelu and P. Blignaut. Convolutional neural network-based methods for eye gaze estimation: A survey. *IEEE Access*, 8:142581–142605, 2020.

[6] K. Alberto Funes Mora and J.-M. Odobez. Geometric generative gaze estimation (g3e) for remote rgb-d cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1773–1780, 2014.

[7] D. Bäck. Neural network gaze tracking using web camera, 2006.

[8] T. Baltrusaitis, P. Robinson, and L.-P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 354–361, 2013.

[9] S. Baluja and D. Pomerleau. Non-intrusive gaze tracking using artificial neural networks. *Advances in Neural Information Processing Systems*, 6, 1993.

[10] J. Bao, B. Liu, and J. Yu. An individual-difference-aware model for cross-person gaze estimation. *IEEE Transactions on Image Processing*, 31:3322–3333, 2022.

[11] Y. Bao, Y. Liu, H. Wang, and F. Lu. Generalizing gaze estimation with rotation consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4207–4216, 2022.

[12] Y. Bao, Y. Liu, H. Wang, and F. Lu. Generalizing gaze estimation with rotation consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4207–4216, 2022.

[13] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing

parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2930–2940, 2013.

[14] P. Biswas et al. Appearance-based gaze estimation using attention and difference mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3143–3152, 2021.

[15] A. Bruce, I. Nourbakhsh, and R. Simmons. The role of expressiveness and attention in human-robot interaction. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, volume 4, pages 4138–4142 vol.4, 2002.

[16] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, pages 1021–1030, 2017.

[17] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE international conference on computer vision*, pages 1513–1520, 2013.

[18] G. T. Buswell. How people look at pictures: a study of the psychology and perception in art. 1935.

[19] M. Cai, F. Lu, and Y. Sato. Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14392–14401, 2020.

[20] J. Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[21] Z. Cao, Z. Chu, D. Liu, and Y. Chen. A vector-based representation to enhance head pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*, pages 1188–1197, 2021.

[22] Z. Cao, Z. Chu, D. Liu, and Y. Chen. A vector-based representation to enhance head pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1188–1197, January 2021.

[23] D. Cazzato, M. Leo, C. Distante, and H. Voos. When i look into your eyes: A survey on computer vision contributions for human gaze estimation and tracking. *Sensors*, 20(13):3739, 2020.

[24] J. J. Cerrolaza, A. Villanueva, and R. Cabeza. Taxonomic study of polynomial regressions applied to the calibration of video-oculographic systems. In *Proceedings of the 2008 symposium on Eye tracking research & applications*, pages 259–266, 2008.

[25] A. K. Chaudhary, R. Kothari, M. Acharya, S. Dangi, N. Nair, R. Bailey, C. Kanan, G. Diaz, and J. B. Pelz. Ritnet: Real-time semantic segmentation

of the eye for gaze tracking. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3698–3702. IEEE, 2019.

[26] L.-C. Chen, M. Collins, Y. Zhu, G. Papandreou, B. Zoph, F. Schroff, H. Adam, and J. Shlens. Searching for efficient multi-scale architectures for dense image prediction. *Advances in neural information processing systems*, 31, 2018.

[27] Z. Chen and B. E. Shi. Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*, pages 309–324. Springer, 2018.

[28] Z. Chen and B. E. Shi. Towards high performance low complexity calibration in appearance based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1174–1188, 2022.

[29] Y. Cheng, Y. Bao, and F. Lu. Puregaze: Purifying gaze feature for generalizable gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 436–443, 2022.

[30] Y. Cheng, S. Huang, F. Wang, C. Qian, and F. Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10623–10630, 2020.

[31] Y. Cheng and F. Lu. Gaze estimation using transformer. *arXiv preprint arXiv:2105.14424*, 2021.

[32] Y. Cheng and F. Lu. Gaze estimation using transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3341–3347. IEEE, 2022.

[33] Y. Cheng, F. Lu, and X. Zhang. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 100–115, 2018.

[34] Y. Cheng, H. Wang, Y. Bao, and F. Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *arXiv preprint arXiv:2104.12668*, 2021.

[35] Y. Cheng, H. Wang, Y. Bao, and F. Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *arXiv preprint arXiv:2104.12668*, 2021.

[36] Y. Cheng, X. Zhang, F. Lu, and Y. Sato. Gaze estimation by exploring two-eye asymmetry. *IEEE Transactions on Image Processing*, 29:5259–5272, 2020.

[37] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[38] E. Chong, N. Ruiz, Y. Wang, Y. Zhang, A. Rozga, and J. M. Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Proceedings of the European conference on*

computer vision (ECCV), pages 383–398, 2018.

[39] V. Clay, P. König, and S. Koenig. Eye tracking in virtual reality. *Journal of eye movement research*, 12(1), 2019.

[40] K. Cortacero, T. Fischer, and Y. Demiris. Rt-bene: A dataset and baselines for real-time blink estimation in natural environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[41] D. Dai, W. Wong, and Z. Chen. Rankpose: Learning generalised feature with rank supervision for head pose estimation. *arXiv preprint arXiv:2005.10984*, 2020.

[42] R. Daza, A. Morales, J. Fierrez, and R. Tolosana. Mebal: A multimodal database for eye blink detection and attention level estimation. In *Companion publication of the 2020 international conference on Multimodal interaction*, pages 32–36, 2020.

[43] F. De la Torre and J. F. Cohn. Facial expression analysis. *Visual analysis of humans: Looking at people*, pages 377–409, 2011.

[44] E. B. Delabarre. A method of recording eye-movements. *The American Journal of Psychology*, 9(4):572–574, 1898.

[45] P. A. Dias, D. Malafronte, H. Medeiros, and F. Odone. Gaze estimation for assisted living environments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 290–299, 2020.

[46] R. Dodge and T. S. Cline. The angle velocity of eye movements. *Psychological Review*, 8(2):145, 1901.

[47] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[48] N. Dubey, S. Ghosh, and A. Dhall. Unsupervised learning of eye gaze representation from the web. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2019.

[49] M. K. Eckstein, B. Guerra-Carrillo, A. T. M. Singley, and S. A. Bunge. Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental cognitive neuroscience*, 25:69–91, 2017.

[50] S. Ellis, R. Candrea, J. Misner, C. S. Craig, C. P. Lankford, and T. E. Hutchinson. Windows to the soul? what eye movements tell us about software usability. In *Proceedings of the usability professionals' association conference*, pages 151–178, 1998.

[51] K. J. Emery, M. Zannoli, J. Warren, L. Xiao, and S. S. Talathi. Openneeds: A dataset of gaze, head, hand, and scene signals during exploration in open-

ended vr environments. In *ACM Symposium on Eye Tracking Research and Applications*, pages 1–7, 2021.

[52] L. Fan, W. Wang, S. Huang, X. Tang, and S.-C. Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5724–5733, 2019.

[53] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. van Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision*, 101(3):437–458, 2013.

[54] T. Fischer, H. J. Chang, and Y. Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 334–352, 2018.

[55] K. A. Funes Mora, F. Monay, and J.-M. Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the symposium on eye tracking research and applications*, pages 255–258, 2014.

[56] S. J. Garbin, Y. Shen, I. Schuetz, R. Cavin, G. Hughes, and S. S. Talathi. Openeds: Open eye dataset. *arXiv preprint arXiv:1905.03702*, 2019.

[57] S. Ghosh, A. Dhall, M. Hayat, J. Knibbe, and Q. Ji. Automatic gaze analysis: A survey of deep learning based approaches. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):61–84, 2023.

[58] S. Ghosh, A. Dhall, M. Hayat, J. Knibbe, and Q. Ji. Automatic gaze analysis: A survey of deep learning based approaches. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):61–84, 2023.

[59] S. Ghosh, A. Dhall, G. Sharma, S. Gupta, and N. Sebe. Speak2label: Using domain knowledge for creating a large scale driver gaze zone estimation dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2896–2905, 2021.

[60] S. Ghosh, M. Hayat, A. Dhall, and J. Knibbe. Mtgls: Multi-task gaze estimation with limited supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3223–3234, 2022.

[61] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021.

[62] E. D. Guestrin and M. Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on biomedical engineering*, 53(6):1124–1133, 2006.

[63] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang,

R. R. Martin, M.-M. Cheng, and S.-M. Hu. Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3):331–368, 2022.

[64] Z. Guo, Z. Yuan, C. Zhang, W. Chi, Y. Ling, and S. Zhang. Domain adaptation gaze estimation by embedding with prediction consistency. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[65] A. Gupta, K. Thakkar, V. Gandhi, and P. Narayanan. Nose, eyes and ears: Head pose estimation by locating facial keypoints. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1977–1981. IEEE, 2019.

[66] D. W. Hansen, J. P. Hansen, M. Nielsen, A. S. Johansen, and M. B. Stegmann. Eye typing using markov and active appearance models. In *Sixth IEEE Workshop on Applications of Computer Vision, 2002.(WACV 2002). Proceedings.*, pages 132–136. IEEE, 2002.

[67] D. W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500, 2009.

[68] D. W. Hansen and A. E. Pece. Eye tracking in the wild. *Computer Vision and Image Understanding*, 98(1):155–181, 2005.

[69] H. Hartridge and L. Thomson. Methods of investigating eye movements. *The British journal of ophthalmology*, 32(9):581, 1948.

[70] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2496–2500. IEEE, 2022.

[71] T. Hempel and A. Al-Hamadi. Slam-based multistate tracking system for mobile human-robot interaction. In *International Conference on Image Analysis and Recognition*, pages 368–376. Springer, 2020.

[72] Q. Hou, D. Zhou, and J. Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13713–13722, 2021.

[73] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[74] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[75] H.-W. Hsu, T.-Y. Wu, S. Wan, W. H. Wong, and C.-Y. Lee. Quatnet: Quaternion-based head pose estimation with multiregression loss. *IEEE Transactions on Multimedia*, 21(4):1035–1046, 2019.

[76] G. Hu, Y. Xiao, Z. Cao, L. Meng, Z. Fang, J. T. Zhou, and J. Yuan. Towards

real-time eyeblink detection in the wild: Dataset, theory and practices. *IEEE Transactions on Information Forensics and Security*, 15:2194–2208, 2019.

[77] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[78] Z. Hu, S. Li, C. Zhang, K. Yi, G. Wang, and D. Manocha. Dgaze: Cnn-based gaze prediction in dynamic scenes. *IEEE transactions on visualization and computer graphics*, 26(5):1902–1911, 2020.

[79] Z. Hu, C. Lv, P. Hang, C. Huang, and Y. Xing. Data-driven estimation of driver attention using calibration-free eye gaze and scene features. *IEEE Transactions on Industrial Electronics*, 69(2):1800–1808, 2021.

[80] B. Huang, R. Chen, W. Xu, and Q. Zhou. Improving head pose estimation using two-stage ensembles with top-k regression. *Image and Vision Computing*, 93:103827, 2020.

[81] B. Huang, R. Chen, W. Xu, and Q. Zhou. Improving head pose estimation using two-stage ensembles with top-k regression. *Image Vis. Comput.*, 93:103827, 2020.

[82] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[83] M. X. Huang, T. C. Kwok, G. Ngai, S. C. Chan, and H. V. Leong. Building a personalized, auto-calibrating eye tracker from user interactions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5169–5179, 2016.

[84] M. X. Huang, T. C. Kwok, G. Ngai, H. V. Leong, and S. C. Chan. Building a self-learning eye gaze model from user interaction data. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1017–1020, 2014.

[85] Q. Huang. *TabletGaze: Dataset and Algorithm for Unconstrained Appearance-based Gaze Estimation in Mobile Tablets.* PhD thesis, Rice University, 2015.

[86] Q. Huang, A. Veeraraghavan, and A. Sabharwal. Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Machine Vision and Applications*, 28:445–461, 2017.

[87] E. B. Huey. Preliminary experiments in the physiology and psychology of reading. *The American Journal of Psychology*, 9(4):575–586, 1898.

[88] E. B. Huey. On the psychology and physiology of reading. i. *The American Journal of Psychology*, 11(3):283–302, 1900.

[89] T. Ishikawa. Passive driver gaze tracking with active appearance models. 2004.

[90] E. Javal. Essai sur la physiologie de la lecture. *Annales d'Oculistique*, 82:242–253, 1879.

[91] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and

Z. Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30:2340–2349, 2021.

[92] S. Jyoti and A. Dhall. Automatic eye gaze estimation using geometric & texture-based networks. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2474–2479. IEEE, 2018.

[93] A. Kar and P. Corcoran. A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. *IEEE Access*, 5:16495–16519, 2017.

[94] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[95] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6912–6921, 2019.

[96] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6912–6921, 2019.

[97] J. Kerr-Gaffney, A. Harrison, and K. Tchanturia. Eye-tracking research in eating disorders: A systematic review. *International Journal of Eating Disorders*, 52(1):3–27, 2019.

[98] D. Keysers, T. Deselaers, C. Gollan, and H. Ney. Deformation models for image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1422–1435, 2007.

[99] A. Khalifa, A. A. Abdelrahman, D. Strazdas, J. Hintz, T. Hempel, and A. Al-Hamadi. Face recognition and tracking framework for human–robot interaction. *Applied Sciences*, 12(11):5568, 2022.

[100] A. Khalifa and A. Al-Hamadi. Jamsface: joint adaptive margins loss for deep face recognition. *Neural Computing and Applications*, pages 1–13, 2023.

[101] J. Kim, M. Stengel, A. Majercik, S. De Mello, D. Dunn, S. Laine, M. McGuire, and D. Luebke. Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12, 2019.

[102] J.-H. Kim and J.-W. Jeong. Gaze estimation in the dark with generative adversarial networks. In *ACM Symposium on Eye Tracking Research and Applications*, pages 1–3, 2020.

[103] A. F. Klaib, N. O. Alsrehin, W. Y. Melhem, H. O. Bashtawi, and A. A. Magableh. Eye tracking algorithms, techniques, tools, and applications with an emphasis on machine learning and internet of things technologies. *Expert Systems with*

Applications, 166:114037, 2021.

[104] Y. Kong and Y. Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

[105] R. Kothari, S. De Mello, U. Iqbal, W. Byeon, S. Park, and J. Kautz. Weakly-supervised physically unconstrained gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9980–9989, 2021.

[106] R. Kothari, Z. Yang, C. Kanan, R. Bailey, J. B. Pelz, and G. J. Diaz. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific reports*, 10(1):2539, 2020.

[107] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016.

[108] A. Kumar, A. Alavi, and R. Chellappa. Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In *2017 12th ieee international conference on automatic face & gesture recognition (fg 2017)*, pages 258–265. IEEE, 2017.

[109] A. Kumar, A. Kaur, and M. Kumar. Face detection techniques: a review. *Artificial Intelligence Review*, 52:927–948, 2019.

[110] J. Lemley, A. Kar, A. Drimbarean, and P. Corcoran. Convolutional neural network implementation for eye-gaze estimation on low-quality consumer imaging systems. *IEEE Transactions on Consumer Electronics*, 65(2):179–187, 2019.

[111] J. L. Levine. *An eye-controlled computer*. IBM Research Division, TJ Watson Research Center, 1981.

[112] S. Li, C. Xu, and M. Xie. A robust o (n) solution to the perspective-n-point problem. *IEEE transactions on pattern analysis and machine intelligence*, 34(7):1444–1450, 2012.

[113] D. Lian, L. Hu, W. Luo, Y. Xu, L. Duan, J. Yu, and S. Gao. Multiview multitask gaze estimation with deep convolutional neural networks. *IEEE transactions on neural networks and learning systems*, 30(10):3010–3023, 2018.

[114] D. Lian, Z. Zhang, W. Luo, L. Hu, M. Wu, Z. Li, J. Yu, and S. Gao. Rgbd based gaze estimation via multi-task cnn. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 2488–2495, 2019.

[115] H. Liu, S. Fang, Z. Zhang, D. Li, K. Lin, and J. Wang. Mfdnet: Collaborative poses perception and matrix fisher distribution for head pose estimation. *IEEE Transactions on Multimedia*, 24:2449–2460, 2021.

[116] Y. Liu, R. Liu, H. Wang, and F. Lu. Generalizing gaze estimation with outlier-guided collaborative adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3835–3844, 2021.

[117] Z. Liu, Z. Chen, J. Bai, S. Li, and S. Lian. Facial pose estimation by deep learning from label distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[118] F. Lu and X. Chen. Person-independent eye gaze prediction from eye images using patch-based features. *Neurocomputing*, 182:10–17, 2016.

[119] F. Lu, T. Okabe, Y. Sugano, and Y. Sato. A head pose-free approach for appearance-based gaze estimation. In *BMVC*, pages 1–11, 2011.

[120] F. Lu, T. Okabe, Y. Sugano, and Y. Sato. Learning gaze biases with head motion for head pose-free gaze estimation. *Image and Vision Computing*, 32(3):169–179, 2014.

[121] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Inferring human gaze from appearance via adaptive linear regression. In *2011 International Conference on Computer Vision*, pages 153–160. IEEE, 2011.

[122] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Adaptive linear regression for appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 36(10):2033–2046, 2014.

[123] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Gaze estimation from eye appearance: A head pose-free method via eye image synthesis. *IEEE Transactions on Image Processing*, 24(11):3680–3693, 2015.

[124] D. Machin. A novel approach to real-time non-intrusive gaze finding. In *Ninth British Machine Vision Conference, 1998*, pages 58–67, 1998.

[125] S. Mehta and M. Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.

[126] S. Mehta and M. Rastegari. Separable self-attention for mobile vision transformers. *arXiv preprint arXiv:2206.02680*, 2022.

[127] X. Mei, Z. Hong, D. Prokhorov, and D. Tao. Robust multitask multiview tracking in videos. *IEEE transactions on neural networks and learning systems*, 26(11):2874–2890, 2015.

[128] M. Meißner and J. Oll. The promise of eye-tracking methodology in organizational research: A taxonomy, review, and future avenues. *Organizational Research Methods*, 22(2):590–617, 2019.

[129] J. Merchant, R. Morrissette, and J. L. Porterfield. Remote measurement of eye direction allowing subject motion over one cubic foot of space. *IEEE transactions on biomedical engineering*, (4):309–317, 1974.

[130] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493:626–646, 2022.

[131] K. A. F. Mora and J.-M. Odobez. Person independent 3d gaze estimation from remote rgb-d cameras. In *2013 IEEE International Conference on Image Pro-*

cessing, pages 2787–2791. IEEE, 2013.

[132] L. Murthy and P. Biswas. Appearance-based gaze estimation using attention and difference mechanism. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3137–3146. IEEE, 2021.

[133] L. Murthy, S. Brahmbhatt, S. Arjun, and P. Biswas. I2dnet-design and real-time evaluation of appearance-based gaze estimation system. *Journal of Eye Movement Research*, 14(4), 2021.

[134] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[135] J. O Oh, H. J. Chang, and S.-I. Choi. Self-attention with convolution and deconvolution for efficient eye gaze estimation from a full face image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4992–5000, 2022.

[136] R. Ogusu and T. Yamanaka. Lpm: learnable pooling module for efficient full-face gaze estimation. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5. IEEE, 2019.

[137] S. Ohayon and E. Rivlin. Robust 3d head tracking using camera pose estimation. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pages 1063–1066. IEEE, 2006.

[138] P. Oza, V. A. Sindagi, V. V. Sharmini, and V. M. Patel. Unsupervised domain adaptation of object detectors: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[139] A. Palazzi, D. Abati, F. Solera, R. Cucchiara, et al. Predicting the driver's focus of attention: the dr (eye) ve project. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1720–1733, 2018.

[140] C. Palmero, J. Selva, M. A. Bagheri, and S. Escalera. Recurrent cnn for 3d gaze estimation using appearance and shape cues. *arXiv preprint arXiv:1805.03064*, 2018.

[141] C. Palmero, A. Sharma, K. Behrendt, K. Krishnakumar, O. V. Komogortsev, and S. S. Talathi. Openeds2020: open eyes dataset. *arXiv preprint arXiv:2005.03876*, 2020.

[142] C. Palmero Cantarino, O. V. Komogortsev, and S. S. Talathi. Benefits of temporal information for appearance-based gaze estimation. In *ACM Symposium on Eye Tracking Research and Applications*, pages 1–5, 2020.

[143] S. Park, E. Aksan, X. Zhang, and O. Hilliges. Towards end-to-end video-based eye-tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 747–763. Springer, 2020.

[144] S. Park, S. D. Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9368–9377, 2019.

[145] S. Park, S. D. Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9368–9377, 2019.

[146] S. Park, A. Spurr, and O. Hilliges. Deep pictorial gaze estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 721–738, 2018.

[147] M. Patacchiola and A. Cangelosi. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, 71:132–143, 2017.

[148] P. Pathirana, S. Senarath, D. Meedeniya, and S. Jayarathna. Eye gaze estimation: A survey on deep learning-based approaches. *Expert Systems with Applications*, 199:116894, 2022.

[149] A. Patney, J. Kim, M. Salvi, A. Kaplanyan, C. Wyman, N. Benty, A. Lefohn, and D. Luebke. Perceptually-based foveated virtual reality. In *ACM SIGGRAPH 2016 emerging technologies*, pages 1–2. 2016.

[150] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics (TOG)*, 35(6):1–12, 2016.

[151] D. Pomerleau and S. Baluja. Non-intrusive gaze tracking using artificial neural networks. In *AAAI Fall Symposium on Machine Learning in Computer Vision, Raleigh, NC*, pages 153–156, 1993.

[152] A. Rangesh, B. Zhang, and M. M. Trivedi. Driver gaze estimation in the real world: Overcoming the eyeglass challenge. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1054–1059. IEEE, 2020.

[153] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:121–135, 2019.

[154] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 17–24, 2017.

[155] G. E. Raptis, C. Katsini, M. Belk, C. Fidas, G. Samaras, and N. Avouris. Using eye gaze data and visual activities to infer human cognitive styles: method and feasibility studies. In *proceedings of the 25th conference on user modeling, Adaptation and Personalization*, pages 164–173, 2017.

[156] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba. Where are they looking?

Advances in neural information processing systems, 28, 2015.

[157] N. Ruiz, E. Chong, and J. M. Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2074–2083, 2018.

[158] N. Ruiz, E. Chong, and J. M. Rehg. Fine-grained head pose estimation without keypoints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[159] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[160] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.

[161] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[162] T. Santini, W. Fuhl, and E. Kasneci. Calibme: Fast and unsupervised eye tracker calibration for gaze-based pervasive human-computer interaction. In *Proceedings of the 2017 chi conference on human factors in computing systems*, pages 2594–2605, 2017.

[163] L. Sesma, A. Villanueva, and R. Cabeza. Evaluation of pupil center-eye corner vector for gaze estimation using a web cam. In *Proceedings of the symposium on eye tracking research and applications*, pages 217–220, 2012.

[164] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 50–58, 2015.

[165] S.-W. Shih, Y.-T. Wu, and J. Liu. A calibration-free gaze tracking technique. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 4, pages 201–204. IEEE, 2000.

[166] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017.

[167] E. Skodras, V. G. Kanas, and N. Fakotakis. On visual gaze tracking based on a single low cost camera. *Signal Processing: Image Communication*, 36:29–42, 2015.

[168] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar. Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the 26th*

annual ACM symposium on User interface software and technology, pages 271–280, 2013.

[169] D. Strazdas, J. Hintz, A. Khalifa, A. A. Abdelrahman, T. Hempel, and A. Al-Hamadi. Robot system assistant (rosa): Towards intuitive multi-modal and multi-device human-robot interaction. *Sensors*, 22(3):923, 2022.

[170] Y. Sugano, Y. Matsushita, and Y. Sato. Appearance-based gaze estimation using visual saliency. *IEEE transactions on pattern analysis and machine intelligence*, 35(2):329–341, 2012.

[171] Y. Sugano, Y. Matsushita, and Y. Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1821–1828, 2014.

[172] Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike. An incremental learning method for unconstrained gaze estimation. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III 10*, pages 656–667. Springer, 2008.

[173] K.-H. Tan, D. J. Kriegman, and N. Ahuja. Appearance-based eye gaze estimation. In *Sixth IEEE Workshop on Applications of Computer Vision, 2002.(WACV 2002). Proceedings.*, pages 191–195. IEEE, 2002.

[174] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[175] M. Tan and Q. Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR, 2021.

[176] A. Tawari, K. H. Chen, and M. M. Trivedi. Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 988–994. IEEE, 2014.

[177] H. Tomas, M. Reyes, R. Dionido, M. Ty, J. Mirando, J. Casimiro, R. Atienza, and R. Guinto. Goo: A dataset for gaze object prediction in retail environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3125–3133, 2021.

[178] M. Tonsen, J. Steil, Y. Sugano, and A. Bulling. Invisibleeye: Mobile eye tracking using multiple low-resolution cameras and learning-based gaze estimation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–21, 2017.

[179] A. Tsukada, M. Shino, M. Devyver, and T. Kanade. Illumination-free gaze estimation method for first-person vision wearable device. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2084–2091. IEEE, 2011.

[180] S. Tulyakov, L. A. Jeni, J. F. Cohn, and N. Sebe. consistent 3d face alignment. *IEEE transactions on pattern analysis and machine intelligence*, 40(9):2250–2264, 2017.

[181] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[182] E. Vahdani and Y. Tian. Deep learning-based action detection in untrimmed videos: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4302–4320, 2022.

[183] R. Valenti and T. Gevers. Accurate eye center location and tracking using isophote curvature. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[184] R. Valenti, J. Staiano, N. Sebe, and T. Gevers. Webcam-based visual gaze estimation. In *Image Analysis and Processing–ICIAP 2009: 15th International Conference Vietri sul Mare, Italy, September 8-11, 2009 Proceedings 15*, pages 662–671. Springer, 2009.

[185] R. Valle, J. M. Buenaposada, and L. Baumela. Multi-task head pose estimation in-the-wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:2874–2881, 2021.

[186] B. Vasli, S. Martin, and M. M. Trivedi. On driver gaze estimation: Explorations and fusion of geometric and data driven approaches. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 655–660. IEEE, 2016.

[187] R. Venkateswarlu et al. Eye gaze estimation from a single image of one eye. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 136–143. IEEE, 2003.

[188] R. Venkateswarlu et al. Eye gaze estimation from a single image of one eye. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 136–143. IEEE, 2003.

[189] H. N. Viet, L. N. Viet, T. N. Dinh, D. T. Minh, and L. T. Quac. Simultaneous face detection and 360 degree head pose estimation. In *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–7. IEEE, 2021.

[190] S. Vora, A. Rangesh, and M. M. Trivedi. Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis. *IEEE Transactions on Intelligent Vehicles*, 3(3):254–265, 2018.

[191] S. N. Wadekar and A. Chaurasia. Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features. *arXiv preprint arXiv:2209.15159*, 2022.

[192] H. Wang, J. Pi, T. Qin, S. Shen, and B. E. Shi. Slam-based localization of 3d gaze using a mobile eye tracker. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, pages 1–5, 2018.

[193] H. Wang and B. E. Shi. Gaze awareness improves collaboration efficiency in a collaborative assembly task. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 1–5, 2019.

[194] H. Wang, J. O. Oh, H. J. Chang, J. H. Na, M. Tae, Z. Zhang, and S.-I. Choi. Gazecaps: Gaze estimation with self-attention-routed capsules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2668–2676, 2023.

[195] K. Wang and Q. Ji. Real time eye gaze tracking with 3d deformable eye-face model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1003–1011, 2017.

[196] K. Wang, H. Su, and Q. Ji. Neuro-inspired eye tracking with eye movement dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9831–9840, 2019.

[197] K. Wang, R. Zhao, H. Su, and Q. Ji. Generalizing eye tracking with bayesian adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11907–11916, 2019.

[198] K. Wang, R. Zhao, H. Su, and Q. Ji. Generalizing eye tracking with bayesian adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11907–11916, 2019.

[199] K. Wang, R. Zhao, H. Su, and Q. Ji. Generalizing eye tracking with bayesian adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11907–11916, 2019.

[200] M. Wang, B. Liu, and H. Foroosh. Design of efficient convolutional layers using single intra-channel convolution, topological subdivisioning and spatial" bottleneck" structure. *arXiv preprint arXiv:1608.04337*, 2016.

[201] Q. Wang, H. Wang, R.-C. Dang, G.-P. Zhu, H.-F. Pi, F. Shic, and B.-l. Hu. Style transformed synthetic images for real world gaze estimation by using residual neural network with embedded personal identities. *Applied Intelligence*, pages 1–16, 2022.

[202] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang. Inferring salient objects from human fixations. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):1913–1927, 2019.

[203] Y. Wang, Y. Jiang, J. Li, B. Ni, W. Dai, C. Li, H. Xiong, and T. Li. Contrastive regression for domain adaptation on gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19376–19385, 2022.

[204] Z. Wang, J. Zhao, C. Lu, F. Yang, H. Huang, Y. Guo, et al. Learning to detect head movement in unconstrained remote gaze estimation in the wild. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3443–3452, 2020.

[205] Z. Wang, Y. Zhao, and F. Lu. Gaze-vergence-controlled see-through vision in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3843–3853, 2022.

[206] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer vision and image understanding*, 115(2):224–241, 2011.

[207] P. Werner, F. Saxen, and A. Al-Hamadi. Landmark based head pose estimation benchmark and method. In *2017 IEEE international conference on image processing (ICIP)*, pages 3909–3913. IEEE, 2017.

[208] O. Williams, A. Blake, and R. Cipolla. Sparse and semi-supervised visual mapping with the s^ 3gp. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 230–237. IEEE, 2006.

[209] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[210] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling. A 3d morphable eye region model for gaze estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 297–313. Springer, 2016.

[211] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 131–138, 2016.

[212] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling. Gazedirector: Fully articulated eye gaze redirection in video. In *Computer Graphics Forum*, volume 37, pages 217–225. Wiley Online Library, 2018.

[213] E. Wood and A. Bulling. Eyetab: Model-based gaze estimation on unmodified tablet computers. In *Proceedings of the symposium on eye tracking research and applications*, pages 207–210, 2014.

[214] H. Wu, Y. Kitagawa, T. Wada, T. Kato, and Q. Chen. Tracking iris contour with a 3d eye-model for gaze estimation. In *Computer Vision–ACCV 2007: 8th Asian Conference on Computer Vision, Tokyo, Japan, November 18-22, 2007, Proceedings, Part I 8*, pages 688–697. Springer, 2007.

[215] Y. Wu, G. Li, Z. Liu, M. Huang, and Y. Wang. Gaze estimation via modulation-

based adaptive network with auxiliary self-learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5510–5520, 2022.

[216] Y. Wu and Q. Ji. Robust facial landmark detection under significant head poses and occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3658–3666, 2015.

[217] Z. Wu, S. Rajendran, T. Van As, V. Badrinarayanan, and A. Rabinovich. Eyenet: A multi-task deep network for off-axis eye gaze estimation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3683–3687. IEEE, 2019.

[218] X. Xiong, Z. Liu, Q. Cai, and Z. Zhang. Eye gaze tracking using an rgbd camera: a comparison with a rgb solution. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 1113–1121, 2014.

[219] Y. Xiong, H. J. Kim, and V. Singh. Mixed effects neural networks (menets) with applications to gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7743–7752, 2019.

[220] M. Xu and F. Lu. Gaze from origin: Learning for generalized gaze estimation by embedding the gaze frontalization process. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6333–6341, 2024.

[221] M. Xu, H. Wang, and F. Lu. Learning a generalized gaze estimator from gaze-consistent feature. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 3027–3035, 2023.

[222] T. Xu, B. Wu, R. Fan, Y. Zhou, and D. Huang. Fr-net: A light-weight fft residual net for gaze estimation. *arXiv preprint arXiv:2305.11875*, 2023.

[223] H. Yamazoe, A. Utsumi, T. Yonezawa, and S. Abe. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. In *Proceedings of the 2008 symposium on Eye tracking research & applications*, pages 245–250, 2008.

[224] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016.

[225] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1087–1096, 2019.

[226] Y. Yang, H. Wei, H. Zhu, D. Yu, H. Xiong, and J. Yang. Exploiting cross-modal prediction and relation consistency for semisupervised image captioning. *IEEE Transactions on Cybernetics*, 2022.

[227] Y. Yang, D.-C. Zhan, Y.-F. Wu, Z.-B. Liu, H. Xiong, and Y. Jiang. Semi-

supervised multi-modal clustering and classification with incomplete modalities. *IEEE Transactions on Knowledge and Data Engineering*, 33(2):682–695, 2019.

[228] A. L. Yarbus and A. L. Yarbus. Eye movements during perception of complex objects. *Eye movements and vision*, pages 171–211, 1967.

[229] D. H. Yoo and M. J. Chung. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *Computer Vision and Image Understanding*, 98(1):25–51, 2005.

[230] Y. Yu, G. Liu, and J.-M. Odobez. Deep multitask gaze estimation with a constrained landmark-gaze model. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[231] Y. Yu, G. Liu, and J.-M. Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11937–11946, 2019.

[232] Y. Yu and J.-M. Odobez. Unsupervised representation learning for gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7314–7324, 2020.

[233] Z. Yu, X. Huang, X. Zhang, H. Shen, Q. Li, W. Deng, J. Tang, Y. Yang, and J. Ye. A multi-modal approach for driver gaze prediction to remove identity bias. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 768–776, 2020.

[234] L. Zeng, L. Chen, W. Bao, Z. Li, Y. Xu, J. Yuan, and N. K. Kalantari. 3d-aware facial landmark detection via multi-view consistent training on synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12747–12758, 2023.

[235] H. Zhang, M. Wang, Y. Liu, and Y. Yuan. Fdn: Feature decoupling network for head pose estimation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12789–12796, 2020.

[236] H. Zhang, M. Wang, Y. Liu, and Y. Yuan. Fdn: Feature decoupling network for head pose estimation. In *AAAI*, 2020.

[237] H. Zhang, Q. Li, and Z. Sun. Adversarial learning semantic volume for 2d/3d face shape regression in the wild. *IEEE Transactions on Image Processing*, 28(9):4526–4540, 2019.

[238] Q. Zhang, Y. Xu, J. Zhang, and D. Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *International Journal of Computer Vision*, pages 1–22, 2023.

[239] W. Zhang, T.-N. Zhang, and S.-J. Chang. Gazing estimation and correction from elliptical features of one iris. In *2010 3rd International Congress on Image and Signal Processing*, volume 4, pages 1647–1652. IEEE, 2010.

[240] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient

convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.

[241] X. Zhang, M. X. Huang, Y. Sugano, and A. Bulling. Training person-specific gaze estimators from user interactions with multiple devices. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

[242] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*, pages 365–381. Springer, 2020.

[243] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 365–381. Springer, 2020.

[244] X. Zhang, Y. Sugano, and A. Bulling. Everyday eye contact detection using unsupervised gaze target discovery. In *Proceedings of the 30th annual ACM symposium on user interface software and technology*, pages 193–203, 2017.

[245] X. Zhang, Y. Sugano, and A. Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, pages 1–9, 2018.

[246] X. Zhang, Y. Sugano, and A. Bulling. Evaluation of appearance-based methods and implications for gaze-based applications. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13, 2019.

[247] X. Zhang, Y. Sugano, A. Bulling, and O. Hilliges. Learning-based region selection for end-to-end gaze estimation. In *31st British Machine Vision Conference (BMVC 2020)*, page 86. British Machine Vision Association, 2020.

[248] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015.

[249] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015.

[250] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–60, 2017.

[251] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 2299–2308. IEEE, 2017.

[252] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017.

[253] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017.

[254] Y. Zhang, A. Bulling, and H. Gellersen. Sideways: A gaze interface for spontaneous interaction with situated displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 851–860, 2013.

[255] Y. Zhang, J. Müller, M. K. Chong, A. Bulling, and H. Gellersen. Gazehorizon: Enabling passers-by to interact with public displays by gaze. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 559–563, 2014.

[256] Y. Zhang, X. Yang, and Z. Ma. Driver's gaze zone estimation method: A four-channel convolutional neural network model. In *2020 2nd International Conference on Big-data Service and Intelligent Computation*, pages 20–24, 2020.

[257] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

[258] Y. Zheng, S. Park, X. Zhang, S. De Mello, and O. Hilliges. Self-learning transformations for improving gaze and head redirection. *Advances in Neural Information Processing Systems*, 33:13127–13138, 2020.

[259] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[260] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 386–391, 2013.

[261] J. Zhou, G. Li, F. Shi, X. Guo, P. Wan, and M. Wang. Em-gaze: eye context correlation and metric learning for gaze estimation. *Visual Computing for Industry, Biomedicine, and Art*, 6(1):8, 2023.

[262] X. Zhou, J. Lin, J. Jiang, and S. Chen. Learning a 3d gaze estimator with improved itracker combined with bidirectional lstm. In *2019 IEEE international conference on Multimedia and expo (ICME)*, pages 850–855. IEEE, 2019.

[263] Y. Zhou and J. Gregson. Whenet: Real-time fine-grained estimation for wide range head pose. *arXiv preprint arXiv:2005.10353*, 2020.

[264] Y. Zhou and J. Gregson. Whenet: Real-time fine-grained estimation for wide range head pose. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

[265] W. Zhu and H. Deng. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3143–3152, 2017.

[266] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, 2012.

[267] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Li. Face alignment across large poses: A 3d solution. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, 2016.

[268] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.

[269] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 787–796, 2015.

[270] Z. Zhu, D. Zhang, C. Chi, M. Li, and D.-J. Lee. A complementary dual-branch network for appearance-based gaze estimation from low-resolution facial image. *IEEE Transactions on Cognitive and Developmental Systems*, 2022.

[271] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023.