Calibration of the Open-Vocabulary Model YOLO-World by Using Temperature Scaling

Max Andreas Ingrisch, Subashkumar Rajanayagam, Ingo Chmielewski and Stefan Twieg
Department of Electrical, Mechanical and Industrial Engineering, HS Anhalt, Bernburger Str. 55, 06366 Köthen, Germany
{*max.ingrisch, subashkumar.rajanayagam, ingo.chmielewski, stefan.twieg*}*@hs-anhalt.de*

Keywords: Calibration, YOLO-World, Temperature Scaling, Expected Calibration Error, Open-Vocabulary Detection.

Abstract: In many areas of the real world, such as robotics and autonomous driving, deep learning models are an indispensable tool for detecting objects in the environment. In recent years, supervised models such as YOLO or Faster R-CNN have been increasingly used for this purpose. One disadvantage of these models is that they can only detect objects within a closed vocabulary. To overcome this limitation, research is currently being conducted into models that can also detect objects outside the known classes of the training data set. A model is therefore trained with base classes and can recognize novel, unseen classes – this is referred to as open-vocabulary detection (OVD). Novel models such as YOLO-World offer a solution to this problem, but they tend to over- or underestimate when calculating confidence values and are therefore often poorly calibrated. However, reliable determination of confidence values is a crucial factor for the use of these models in the real world to ensure safety and trustworthiness. To address this problem, this paper investigates the influence of the calibration method temperature scaling on the OVD model YOLO-World. The optimal T-value is determined by 2 calibration data sets (Pascal VOC and Open Images V7) and then evaluated on the LVIS minival dataset. The results show that the use of temperature scaling improved the Expected Calibration Error (ECE) from 6.78% to 2.31%, but the model still tends to overestimate the confidence values in some bins.

1 INTRODUCTION

Object detection plays an essential role in many areas and systems. These include robotics and autonomous driving, where it enables interaction with the environment (e.g. traffic sign recognition [1]) and helps to avoid unwanted collisions. In recent years, deep learning (DL) techniques have been used to develop models such as *YOLO* [2] and *Faster R-CNN* [3]. However, these models are limited to the detection of objects that were learned during training (closed-set object detection). To overcome this limitation, more research has been done on models that can also detect objects outside the training data set. This is called Open-Vocabulary Detection (OVD). [4] CLIP (Contrastive Language-Image Pre-training) [5], which involves the joint training of a text and image encoder, is an important step towards the realization of an OVD model. This approach attempts to extract text features and image features and map them in a common embedding space. Based on CLIP, OVD models such as *YOLO-World* [6] or *Det-CLIP* [7] were developed. There are a variety of other OVD models such as *Grounding DINO* [8], which uses a BERT text encoder instead of CLIP.

Despite the achieved generalization capability of these models, there is currently a problem that the calculated confidence values are not always reliable, as shown for Grounding DINO in [9]. This means that the models tend to be over- or underconfident – in some cases, true positives (TP) have too low confidence values, while false positives (FP) have too high. The models are therefore not well calibrated. Without a reliable determination of the confidence values, the use of such models in the real world is associated with some risks - for example, if a red traffic light is identified as green with 95% certainty in 90% of cases.

To overcome the problem of miss-calibration, there are numerous methods, which are presented in [10] and [11] for Neural Networks. In this paper, the calibration of YOLO-World is first examined by calculating the Expected Calibration Error (ECE) and displaying the corresponding Reliability Diagram. From this, it can be deduced how large the calibration error is and whether the model tends to over- or underestimate the confidence values. Subsequently, the simple calibration method *Temperature Scaling* is applied in accordance with [12] and the

influence on the confidence values and the calibration error is determined. Two calibration data sets, Pascal VOC [13] and Open Images V7 [14, 15], are used to determine the optimal T-value. Furthermore, the influence of the calibration on the accuracy of the model is examined by calculating the mAP (mean Average Precision) for each determined T-value. YOLO-World was selected for this study because, according to [6], it outperforms other models, such as Grounding DINO and Det-CLIP, in terms of mAP (mean Average Precision) on the LVIS minival [16] dataset.

2 BACKGROUND

2.1 YOLO-World

YOLO-World [6] is an OVD model that can detect novel classes, i.e., objects that were not included in the training data set, in an image. For this purpose, YOLO-World uses a frozen text encoder (CLIP), which generates the text embeddings $W = TextEncoder(F) \in \mathbb{R}^{C \times D}$ from a given text F, where C is the number of classes and D is the dimensionality of the embeddings. Here, $w_j \in W$ represents the j-th text embedding. The YOLO backbone (YOLOv8) then extracts the image features of an image I. From the given input of image I and text F, K object embeddings $\{e_k\}_{k=1}^{K}$ with $e_k \in \mathbb{R}^D$ are then generated. The logit (non-probabilistic output of the network [11]) can thus be formulated as

$$l_{k,j} = \alpha \cdot \text{Batch-Norm}(e_k) \cdot \text{L2-Norm}(w_j)^T + \beta \,. \quad (1)$$

In equation 1, $l_{k,j}$ represents the logit, i.e., the object-text similarity, between the k-th object embedding and the j-th text embedding. The text and object embeddings are previously L2-normalized or Batch-normalized and the product is scaled by two constants α and β, which are learned during the training process. After that, the logits are activated by a sigmoid function [17], which is defined as

$$\sigma(l_{k,j}) = \frac{1}{1 + e^{-l_{k,j}}} \,. \quad (2)$$

In a sigmoid function, the probability $\sigma(l_{k,j})$ of each logit is calculated independently, in contrast to the softmax function.

2.2 Temperature Scaling

The *temperature scaling* calibration method is a method that is applied after the training process of an DL model. For this, the logits are scaled with a temperature value $T > 0$, i.e. a constant value. [10, 11] The optimal temperature value T must be learned using a separate calibration data set. According to [10, 11], this method is very efficient and performs better than other calibration methods in vision tasks. The scaled logits can be calculated using the equation

$$l_{k,j}^{cal} = \frac{l_{k,j}}{T} \,, \quad (3)$$

where $l_{k,j}^{cal}$ are the logits after applying temperature scaling.

2.3 Mean Average Precision (mAP)

The mAP (mean Average Precision) is a metric for determining the performance of a model. To calculate this metric, the AP (average precision) value is averaged across all classes K [18]:

$$mAP = \frac{1}{K} \sum_{i=1}^{K} AP_i \,. \quad (4)$$

To determine the AP, two further values, *recall* and *precision*, must first be determined, which can be derived from the number of true positives (TP), false positives (FP) and false negatives (FN). Further information can be found in [19]. For each confidence threshold, a recall and a precision value is calculated and displayed in a so-called *recall-precision curve*. The approximated area under this curve then yields the AP.

2.4 Reliability Diagram and Expected Calibration Error (ECE)

Reliability diagrams can be used to visualize the calibration of a model graphically. [10, 11] Let i be a sample consisting of the maximum confidence value $\hat{p}_i$, the associated predicted label $\hat{y}_i$, and the true class label y_i. First, all samples i are divided into M bins $(B_1, ..., B_M)$ based on their confidence values $\hat{p}_i$, where the interval size is $1/M$. B_m contains the set of samples that fall within the confidence interval $I_m = (\frac{m-1}{M}, \frac{m}{M}]$, where $m \in 1, ..., M$. According to this classification, two metrics can be derived that are essential for the visualization of the reliability diagram as well as for the calculation of the expected calibration error. The accuracy $acc(B_m)$, which is defined as

$$acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i) \,, \quad (5)$$

determines the ratio of the number of samples that have correctly predicted the label to the total number of samples ($|B_m|$) of the bin B_m. In contrast to

accuracy, the confidence of B_m,

$$conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \,, \tag{6}$$

calculates the mean of all predicted confidence values $\hat{p}_i$ in the bin B_m.

In a reliability diagram, the accuracy and confidence are then displayed for each bin B_m, where a deviation from the diagonal $f(x) = x$ symbolizes a miss-calibration. [10, 12] In Figure 2, a blue bar indicates the $acc(B_m)$, while a red bar always indicates a gap between $acc(B_m)$ and $conf(B_m)$, i.e., a calibration error. For a perfectly calibrated model is $acc(B_m) = conf(B_m)$ for all $m \in 1,...,M$. Since the reliability diagram does not take into account the number of samples per bin ($|B_m|$), a confidence histogram is also illustrated. This provides an overview of the percentage of samples, based on the total number, in each bin B_m (see Figure 2).

The Expected Calibration Error (ECE) reflects a value that quantitatively captures the calibration of the model. To do this, the weighted mean of the difference between $acc(B_m)$ and $conf(B_m)$ is calculated,

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \mathrm{acc}(B_m) - conf(B_m) \right| . \tag{7}$$

Here, n is the total number of samples. For a perfectly calibrated model, $ECE = 0$.

3 METHODOLOGY

The aim of this paper is to examine the calibration of YOLO-World and to analyze the influence of the calibration method *temperature scaling* while maintaining the accuracy and performance of this model (mAP).

The YOLO-World model used in this study is the *YOLO-Worldv2-L* model (size: 1280), which was pretrained using the *Objects365* [20] and *GoldG* [16] datasets (see [17]). A maximum of 300 detections are considered per image. For the evaluation, the LVIS minival [16] dataset (a subset of LVIS [21]) with 5000 images and 1203 object classes is used for all investigations. Furthermore, the localization of the objects is done using bounding boxes.

For the implementation, the ECE and the reliability diagram of the non-calibrated YOLO-World are calculated first. The number of bins is $M = 10$ for all the investigations carried out in this work. For both the ECE and the reliability diagram, the true positives (TP) and false positives (FP) must be determined. The basic process is shown in Figure 1 and consists of four steps. In the first step, all ground truths (GT) of an image are compared with each prediction, consisting of a label, a confidence score and the bounding box. Initially, all matches are considered that the correct label predicted and where the calculated IoU (Intersection over Union) value between GT and prediction is above an IoU threshold. After that, they are sorted in descending order based on their score (second step). The matches with the highest score are selected first and evaluated as TP, whereby each prediction may be assigned to a maximum of one ground truth and each ground truth to a maximum of one prediction. In this way, the matches with the highest scores are preferred and those with a lower score are evaluated later. This is the third step. The remaining matches and predictions are categorized as false positives (step four). As a result, the true positives with the best scores and the false positives can be determined, which are essential for calculating the ECE (see equation 7). In this work, the two thresholds $IoU\text{-}Threshold \in \{0.5, 0.75\}$ are considered, which are common threshold values for determining the mAP (see [6]).

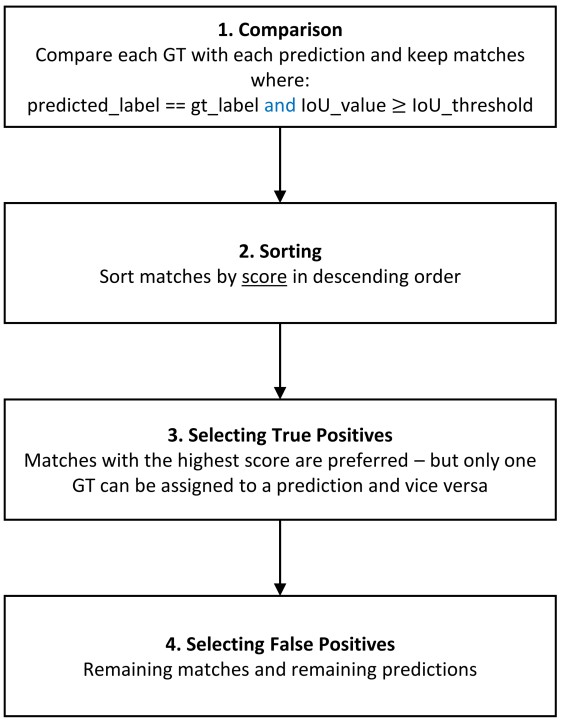

Figure 1: Process for determining the True Positives and False Positives

Subsequently, two calibration data sets, Pascal VOC [13] and Open Images V7 [14, 15], are used to determine the optimal temperature value. Of the Pascal VOC data set, only the validation data set is used, which includes 20 object classes and 5823 images. Since the Open Images validation dataset contains over 40,000 images, only a subset (8000 images)

is used. These can be extracted in a deterministic way using this Python code, whereby the total number of object classes (601 in total) is not changed:

```
import fiftyone as fo
fo.zoo.load_zoo_dataset(
   "open-images-v7",
   split="validation",
   label_types=["detections"],
   max_samples=8000)
```

In accordance with [12], it is determined that T can take on the values $T \in \{0.5 + 0.05 \cdot n \mid n \in \mathbb{N}, 0 \leq n \leq 30\}$. The distance of 0.05 was chosen for reasons of time and resources. For each of these T-values, the ECE of the two calibration data sets can then be determined and displayed graphically, with the scaling according to equation 3. The minima (minimum T-values) obtained in this way are then evaluated by the LVIS minival data set. This makes it possible to see whether the found minima also reduce the ECE of the LVIS minival evaluation data set and whether there is a correlation between the minima of the data sets. By comparing the ECE values and the reliability diagrams before and after scaling with T, the effectiveness of the *temperature scaling* procedure with regard to the calibration of the YOLO-World model can be derived.

4 RESULTS

4.1 Before Applying Temperature Scaling

As described in Section 3, the calibration of the YOLO-World model is first examined using the LVIS minival evaluation dataset. The results in Table 1 show the values of the ECE and the mAP for different IoU thresholds before *temperature scaling*. The ECE has a size of 6.78% and 7.09% at IoU thresholds of 0.5 and 0.75, respectively.

Table 1: ECE and mAP for various IoU thresholds before applying temperature scaling.

Metric	Result [%]
mAP@[IoU=0.5:0.95]	34.6
mAP@[IoU=0.5]	45.5
mAP@[IoU=0.75]	37.8
ECE@[IoU=0.5]	6.78
ECE@[IoU=0.75]	7.09

Looking at the reliability diagram and the confidence diagram (see Figure 2), it can be seen that about 75% of the predictions lie in the confidence range of $(0;0.1]$. The majority of the predictions therefore have only a very low score (confidence value). Furthermore, there is a deviation from the diagonal in all bins. The accuracy ($acc(B_m)$), shown in blue, differs significantly from the calculated confidence ($conf(B_m)$), which can be seen from the red gap. This indicates that the model tends to be overconfident in its predictions in all bins.

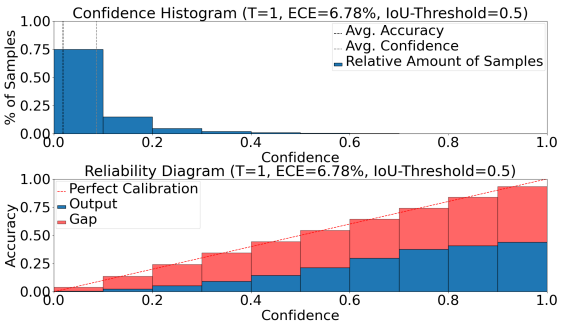

Figure 2: Confidence histogram and reliability diagram for an IoU threshold of 0.5 and $T = 1$.

4.2 After Applying Temperature Scaling

To determine the optimal T-value, the *temperature scaling* methode is applied to the calibration data sets as described in section 3. At this point, it should be mentioned that calibration methods such as temperature scaling do not model either the data or the model uncertainty of a model. [10] The results of this investigation are illustrated in Figure 3. As it can be seen, there are a total of 4 optimal T-values. The temperature values of the two IoU thresholds (0.5 and 0.75, respectively) of the Open Images calibration dataset are relatively close to each other ($T = 0.65$ and $T = 0.6$, respectively), while there are larger deviations for the Pascal VOC dataset ($T = 0.85$ and $T = 1.2$, respectively).

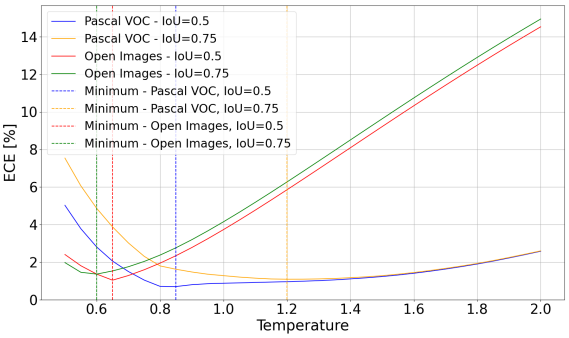

Figure 3: A plot of the ECE as a function of T for the two calibration data sets, Pascal VOC and Open Images, to determine the minima for the IoU thresholds 0.5 and 0.75.

The T-scores obtained in this way are then applied to the LVIS minival evaluation dataset. The accuracy and performance of the model are retained after application, i.e. the mAP remains unchanged for all T-values and IoU thresholds, as listed in Table 1. However, Table 2 shows that the ECE is significantly reduced, which assumes the most minimal value of $ECE = 2.31\%$ and $ECE = 2.66\%$ at $T = 0.6$, respectively. In general, the temperature values determined by Open Images result in the lowest ECE. By using the Pascal VOC, a T-value ($T = 0.85$) could be determined, which also reduces the ECE, but also a T-value ($T = 1.2$), which leads to an increase in ECE. From this, it can be deduced that the Open Images dataset is best suited for calibration using temperature scaling under the given experimental conditions. One possible reason could be the large number of 601 object classes, whereas Pascal VOC comprises just 20, which are more or less covered by the pre-training dataset *Objects365* of the YOLO-World model. A temperature value of $T = 1$ represents the results without applying the calibration method. Thus, temperature scaling could reduce the calibration error ECE by 4.47% and 4.43% for an IoU threshold of 0.5 and 0.75, respectively, in the best case ($T = 0.6$).

Table 2: Determined ECE values for various IoU thresholds after applying temperature scaling.

T value	IoU threshold	ECE [%]
0.6	**0.5**	**2.31**
0.6	**0.75**	**2.66**
0.65	0.5	2.70
0.65	0.75	3.03
0.85	0.5	4.84
0.85	0.75	5.16
1	0.5	6.78
1	0.75	7.09
1.2	0.5	9.45
1.2	0.75	9.76

However, looking at the confidence or reliability diagram in Figure 4 again after calibration, the first thing to note is that now about 90% of the samples lie in the confidence interval of $(0;0.1]$. The two diagrams are shown here for $T = 0.6$ and an IoU threshold of 0.5. After applying temperature scaling, a significant improvement in calibration within bin B_1 can be seen, as well as a slight improvement in bins B_2 to B_4. However, the model still tends to be overconfident in its predictions, especially in bins B_3 to B_{10}, despite a reduction in ECE. It should be noted that the ECE is calculated from a weighted sum (see equation 7). Since about 90% of the samples are in the first bin, this also has the greatest influence on the calculation of the ECE. In summary, it can be deduced from this study that the proposed method could reduce the ECE, but without reducing the overconfidence of the model in all bins.

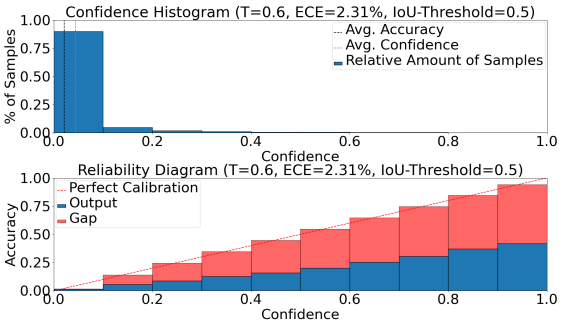

Figure 4: Confidence histogram and reliability diagram for an IoU threshold of 0.5 and $T = 0.6$.

5 CONCLUSION

In this paper, the influence of the calibration method *temperature scaling* on the OVD model YOLO-World was investigated. The optimal T-value could be determined using two calibration data sets, with the Open Images data set proving to be more suitable. It was shown that the model without the application of temperature scaling has a calibration error (ECE) of 6.78% and 7.0% for an IoU threshold of 0.5 and 0.75, respectively, and is overconfident in its predictions. After application, the calibration error ECE is reduced to a value of 2.31% (or 2.66%), while maintaining the accuracy and performance of the model. However, the overconfidence of the model could only be reduced in a few bins in this way. Additional or different calibration approaches are necessary to further improve the calibration of the model and thus minimize overconfidence in the predictions. An accurate modeling of the model uncertainty could also improve the calibration.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the Research, Transfer and Start-Up Center (Forschungs-, Transfer- und Gründerzentrum), the Anhalt University of Applied Sciences and the state of Saxony-Anhalt for our research.

REFERENCES

[1] S. Twieg and R. Menghani, "Analysis and implementation of an efficient traffic sign recognition based on YOLO and SIFT for TurtleBot3 robot," 2023, [Online]. Available: http://dx.doi.org/10.25673/112993.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 2017.

[4] C. Zhu and L. Chen, "A survey on open-vocabulary detection and segmentation: Past, present, and future," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 12, pp. 8954-8975, 2024.

[5] A. Radford et al., "Learning transferable visual models from natural language supervision," 2021, [Online]. Available: https://arxiv.org/abs/2103.00020.

[6] T. Cheng et al., "YOLO-World: Real-time open-vocabulary object detection," in 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 16901-16911.

[7] L. Yao et al., "DetCLIP: Dictionary-enriched visual-concept paralleled pretraining for open-world detection," in Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 9125-9138, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/3ba960559212691be13fa81d9e5e0047-Paper-Conference.pdf.

[8] S. Liu et al., "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," 2024, [Online]. Available: https://arxiv.org/abs/2303.05499.

[9] F. Mumuni and A. Mumuni, "Segment anything model for automated image data annotation: Empirical studies using text prompts from Grounding DINO," 2024, [Online]. Available: https://arxiv.org/abs/2406.19057.

[10] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," 2022, [Online]. Available: https://arxiv.org/abs/2107.03342.

[11] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," 2017, [Online]. Available: https://arxiv.org/abs/1706.04599.

[12] W. LeVine et al., "Enabling calibration in the zero-shot inference of large vision-language models," 2023, [Online]. Available: https://arxiv.org/abs/2303.12748.

[13] M. Everingham et al., "The PASCAL Visual Object Classes challenge: A retrospective," International Journal of Computer Vision, vol. 111, no. 1, pp. 98-136, 2015.

[14] A. Kuznetsova et al., "The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale," arXiv:1811.00982, 2018.

[15] I. Krasin et al., "OpenImages: A public dataset for large-scale multi-label and multi-class image classification," 2017, [Online]. Available: https://storage.googleapis.com/openimages/web/index.html. [Accessed: 2-Jan-2025].

[16] A. Kamath et al., "MDETR - Modulated detection for end-to-end multi-modal understanding," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1760-1770.

[17] Tencent AI Lab, "YOLO-World GitHub Repository," [Online]. Available: https://github.com/AILab-CVC/YOLO-World. [Accessed: 2-Jan-2025].

[18] T. Hirsch and B. Hofer, "The map metric in information retrieval fault localization," in 2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE), 2023, pp. 1480-1491.

[19] P. Zhang and W. Su, "Statistical inference on recall, precision and average precision under random selection," in 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, 2012, pp. 1348-1352.

[20] S. Shao et al., "Objects365: A large-scale, high-quality dataset for object detection," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8429-8438.

[21] A. Gupta, P. Dollar, and R. Girshick, "LVIS: A dataset for large vocabulary instance segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.