Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.38494

Open-Vocabulary Object Detection Based on Context-Awareness and Hierarchical Classification

Haoze Li

School of Digital Science, Shanghai Lida University, Shanghai, China
accident314284@gmail.com

Abstract. In real-world open-scene scenarios, object detection often suffers from bottlenecks such as uneven distribution of tail classes, failure to detect small objects and corruption due to background noise and that is why traditional closed domain detection algorithms will behave badly when presented with new or unlabeled classes. The paper proposes an object detection architecture that can support open vocabulary conditions, named APOVD-Net (Adaptive Prompt Open Vocabulary Detection Network). To address the issue, the network adopts a hierarchical classifier, which follows a coarse-to-fine, large-to-small framework and uses hierarchical semantics routing to provide finer disambiguation. Additionally, a dual-confidence adaptive evaluation is suggested to get a dynamic trade-off in the spatial localization accuracy and cross-modal semantic agreement. Experiments conducted on the official LVIS v1.0 validation set show that APOVD-Net outperforms in both average precision (mAP) and rare category detection rate (Rare AP), although with a very high inference rate, and can be regarded as a viable intelligent visual perception framework for open environments.

Keywords: open-vocabulary object detection, context-aware fusion, hierarchical classifier, dual confidence assessment, vision-language model

1. Introduction

Object recognition is one of the fundamental tasks in computer vision and plays an important role in real-life applications, including intelligent security, self-driving, and robot perception [1, 2].

Traditional closed-set object detection methods (e.g., Faster R-CNN, YOLO series) are based on pre-defined discrete category labels and massive manually annotated datasets [3]. In practice, visual scenes may include untrained objects, long-tail distributions, and new categories. The conventional methods have a number of issues in effectively extrapolating and identifying objects not in the training set, including failure to identify them or lack of sufficient strength [4]. Open-Vocabulary Object Detection (OVD) was created to address the shortcomings of closed-set labeling [5].

By leveraging the extensive image-text alignment capability of contrastive language-image pretraining models (e.g., With CLIP), OVD allows models to take any text description as an object detection target [6, 7]. Nevertheless, current open-vocabulary object detection systems are still faced with two significant issues in real-world implementation. One is the absence of contextual environmental information. Because visual features are directly obtained on the cropped regions

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.38494

produced by the detector, feature encoding may be distorted when dealing with densely occluded or very small targets [8]. Another is the existence of fine-grained semantic ambiguity.

When dealing with complicated scenarios that have thousands of categories covering extensive vocabularies (the LVIS dataset), matching objects based solely on cosine similarity can easily lead to false positives between fine-grained categories that have a similar shape (such as different types of birds or tools) [9]. Simultaneously, the mismatch between localization prediction and semantics scoring is also a significant factor in driving the high false-positive rates. Since there is no plausible joint calibration of the object presence score provided by Region Proposal Network (RPN/YOLO) and the semantic alignment score provided by the visual-language model, the outcomes of classifications are biased [10]. To overcome these bottlenecks, this paper proposes an open-vocabulary detection system called APOVD-Net. It has three components. Firstly, it has a context sensitive feature fusion module at the C-ROI step, which takes advantage of Context Crop technology to learn environmental semantics and incorporates a channel attention mechanism, which can be applied to adaptively complementing object features and contextual information. Secondly, in order to reduce ambiguity due to long-tail distribution, it makes use of a two-stage hierarchical classifier, namely, coarse-grained routing-fine-grained disambiguation. Finally, it proposes a dual-confidence adaptive calibration algorithm, which has proven to be extremely effective at combining localization and classification confidence scores, reducing the false-positive rate significantly.

2. Data and algorithm

2.1. Dataset

This study evaluates the proposed method by means of the publicly available LVIS v1.0 high-difficulty benchmark dataset [11]. This dataset may be considered as an authoritative testbed of open-set and long-tail object recognition problems consisting of 1,203 high-quality categories of objects covering all frequency ranges, i.e., rare, frequent, and common objects. The experimental analysis involves 5,000 high-resolution images that have been selected carefully based on the Val2017 dataset and served as a validation set. Preprocessing is done before the images are supplied to the network to improve the quality of the image under low light or high contrast conditions by applying the Adaptive Limited Contrast Stretching Linear Expansion (CLAHE) algorithm. To enhance the generalization capacity of the models as well as their strength in complex lighting conditions and various viewpoints, the preprocessing pipeline also adds data augmentation techniques such as proportional scaling and random horizontal flipping, used for training. Furthermore, the comparative experiments are conducted in the identical hardware environment, training and evaluation parameters, so that the performance comparison among various algorithms can be made objective and rigorous.

2.2. Network architecture

2.2.1. C-ROI context-aware feature fusion module

Considering the candidate boxes B given by the detector, which represent candidate object regions. The first step is to extract it. To achieve the goal of adding environmental semantic information that does not lie within the limit, this paper proposes a mechanism of extended candidate box:

DOI: 10.54254/2755-2721/2026.38494

$$B_{ctx} = \exp and(B, \alpha) \tag{1}$$

Here the adaptive boundary extension factor σ is chosen to be 1.3.

Subsequently, the cropped image block is used as the input to OpenCLIP visual encoder in order to produce both subject feature vectors and context feature vectors [12], which are subsequently delivered to the adaptive feature fusion module to achieve collaborative multi-scale feature mapping [13]:

$$V_{fused} = \sigma\left(W_1 * V_{ctx}\right) \odot V_{crop} + V_{dx} \tag{2}$$

In this case, both W_1 and V are learnable weight matrices; $\sigma(.)$ denotes the Sigmoid activation function; and $\odot$ denotes the Hadamard product (element-wise multiplication). On this combination strategy, the model may be able to correctly identify the properties of the target object as well as have the liberty to consider surrounding context information. It is important to note that there is no use of simple feature concatenation/sum in the present paper but an adaptive fusion method based on a gating mechanism using the learnable weight matrices is proposed, which dynamically adjusts the contribution of wo types of features. The weights of the gating derived by the application of the Sigmoid activation function allow regulating the level of integration of contextual semantics in real time in relation to the intensity of the occlusion of the target, its size and complexity of the environment. In case of severe occlusion of the target or in case of extremely small objects, the gating mechanism can automatically assign higher weights to contextual features to compensate the absence of primary characteristics of the target with environmental semantics; vice versa, when the target is clearly visible and not occluded, the model focuses on highlighting the primary characteristics of the target and completely suppressing the impact of irrelevant background noise. This adaptive fusion device is much more efficient in solving the issue of feature encoding distortion at higher density occlusions or very small objects, providing a stronger feature representation at the next hierarchy classification.

2.2.2. Hierarchical semantic classifier

The solution proposed in this paper involves the mapping of the 1,203 categories of the LVIS dataset to a tree-like topology of superclass-subclass relationships and hence constructing a semantic network based on hierarchies. The disambiguation process has two steps in the fused feature vector case. First, the similarity is calculate between the combined features and the text embedding representations of every superclass (parent class) to perform coarse-grained semantic routing and classification. Then select the most likely K parent-class branches with the highest predicted probabilities as potential routes to the next level of fine-grained classification:

$$P\left(c_i v_{fused}\right) = \frac{\frac{\exp\left(\cos\left(v_{fused}, T_i\right)\right)}{\tau}}{\sum_{j \in C_{super}} \frac{\exp\left(\cos\left(v_{fused}, T_i\right)\right)}{\tau}} \tag{3}$$

DOI: 10.54254/2755-2721/2026.38494

In summary, the final step of fine-scale disambiguation is done by re-matching in the chosen parent class branch with the corresponding subclass text embeddings in order to get the final predicted category label.

2.2.3. Dual-confidence adaptive evaluation

Detector localization score (YOLO) and hierarchical classifier semantic correspondence score are not on the same scale [14]. The proposed dual-confidence joint formula is adaptive in this study:

$$S_{fina} = (S_{det})^{\gamma} \times (S_{cls})^{1-\gamma} \times (1 - \gamma \bullet \frac{Area(B)}{Area(I_{img})}) \tag{4}$$

where γ is the balance term, and the hyperparameter, which is used to penalize big bounding boxes is l. Lastly, the Soft-NMS algorithm is used to decrease and filter redundant bounding boxes which results in the final detection.

3. Experimental results

3.1. Evaluation metrics

The test is based on the official LVIS assessment protocol, which uses mAP as the primary measure. The evaluation also consists of all average precision mAP (All), rare category mAP (Rare), common category mAP (Common), frequent category mAP (Frequent), and average inference frame rate FPS (Frames Per Second).

The mAP (All), which measures the average precision of all categories, is used as the main measure of the overall performance of the detectors. The mAP (Rare) measures the average precision of the rare categories only, providing a less complicated picture of the model performance on the low-frequency classes in long-tailed distributions. The mAP (Common) and mAP (Frequent) are the indicators of the model performance on common and frequent categories respectively and are used to analyze how performance is distributed at different levels of frequency. FPS is an indicator of model inference speed, useful to test the real-time capabilities of a model in practice. This study will incorporate all these metrics into a single dataset and evaluation framework to reflect the models performance in a comprehensive and objective manner.

3.2. Algorithmic network architecture

In order to test the effectiveness of APOVD-Net, the present study compares it with other popular methods of open-vocabulary detection on the LVIS v1.0 validation dataset (Table 1).

Table 1. Quantitative comparison outcomes on the LVIS v1.0 validation dataset

Method	Backbone	mAP(All) (%)	mAP(Rare) (%)	mAP(Common)(%)	mAP(Frequent)(%)	Frame rate(FPS)
Faster R-CNN [3]	ResNet-50	22.5	11.2	21.4	28.1	18.5
ViLD [6]	ResNet-50	25.5	16.6	24.8	30.3	8.2
Detic [8]	Swin-B	32.4	24.9	32.1	36.8	11.4

DOI: 10.54254/2755-2721/2026.38494

Table 1. (continued)

APOVD-Net	YOLOv8 +CLIP-ViT-B/16	34.8	28.6	34.1	38.5	24.6

The experimental findings presented in Table 1 show that the suggested APOVD-Net performs with an mAP All of 34.8%, which represents an improvement of 9.3 percentage points over the conventional ViLD approach.

In the Rare mAP, a crucial measure of assessing the quality of open-vocabulary generalization, the APOVD-Net scores an impressive 28.6%, which is much higher than the baseline models. This result further demonstrates the spectacular efficiency of the C-ROI contextual fusion and hierarchical classifier to address the bottlenecks of rare and few-shot object recognition. Further analysis of comparative data in Table 1 will indicate that our method has more balanced benefits in the performance of all evaluation dimensions. APOVD-Net outperforms Faster R-CNN and ViLD by 12.3 and 9.3 percentage points respectively when ResNet-50 is used as the backbone network on mAP All. Even compared to the Detic model, which uses the more potent Swin-B backbone network, our approach outperforms 2.4 percentage points. Rare mAP shows 28.6% score, outperforming ViLD and Detic by 12.0 and 3.7 percentage points respectively, confirming conclusively that the contextual reconstruction and hierarchical disambiguation at the C-ROI stage effectively improve the feature representation ability of rare classes significantly [15]. Moreover, due to the synergy of YOLOv8 lightweight architecture and batch processing optimization, APOVD-Net has a real-time inference rate of 24.6 FPS with the highest detection accuracy, far above those of mainstream competing algorithms. This is indicative of a great combination of both accuracy and speed and this model has a lot of practical value when it comes to its application in the real world context.

3.3. Ablation study

In order to investigate the role of each core module in improving network performance thoroughly, this paper conducts ablation experiments which are shown in Table 2.

Table 2. Ablation study results of the core modules in APOVD-Net

Experiment Number	C-ROI Contextual Fusion	Hierarchical classifier	Dual Confidence Calibration	mAP(All) (%)	mAP(Rare) (%)
1	T	F	F	26.1	18.2
2	T	F	F	29.4	22.5
3	T	T	F	32.8	26.1
4	T	T	T	34.8	28.6

As per Table 2, the outcomes of the ablation show that each of the core modules is positively contributing to the network performance. The values of mAP (All) and mAP (Rare) are 26.1% and 18.2% respectively in the foundational setup (Experiment 1) which only uses the basic detection framework. After introducing the C-ROI context fusion module (Experiment 2), these numbers have increased to 29.4% and 22.5% respectively indicating the importance of environmental context information in improving feature robustness. The addition of the hierarchical classifier (Experiment 3) raises mAP (All) and mAP (Rare) to 32.8 and 26.1, respectively. It demonstrates that the two-stage classification between coarse and fine categories is effective in reducing the problem of class

DOI: 10.54254/2755-2721/2026.38494

confusion in the case of large vocabularies and, in particular, improves recognition performance of rare categories. The highest results of the model are obtained (Experiment 4) when all the three modules are fully activated with mAP (All) and mAP (Rare) at 34.8 and 28.6 respectively, confirming that the three modules work together and are important

4. Conclusion

This paper investigates the main bottlenecks of open-vocabulary object recognition, including the imbalance of long-tail distributions, lack of contextual information, and mismatch between localisation and classification score. Through a context-sensitive feature adaptive fusion process in the C-ROI module, APOVD-Net is significantly more able to depict small or occluded objects in cluttered scenes. Moreover, the hierarchical classifier and a dual-similarity reduction algorithm, it can decrease semantic ambiguity between different categories and improves the effectiveness of post-processing filtering. Experiments on the LVIS benchmark set show that APOVD-Net is very effective in terms of overall accuracy, ability to generalize to rare classes, and inference rate. In the present paper, four aspects are explored in detail. Firstly, future work will extend the existing two-dimensional static detection structures to accommodate more complex modalities like dynamic video streams and panoramic images, and therefore validating the constraints of the generalization in dynamic spatiotemporal environment. Secondly, it will employ more powerful vision-language large language models to investigate the mechanisms of open-vocabulary representation learning further and hence achieve a better fine-grained semantic disambiguation capability. Thirdly, it will implement light-weight methods such as model quantization and knowledge distillation to allow edge deployment, i.e., reduce the size of the model and its computational cost without loss of accuracy. Finally, it will enhance the size and diversity of training vocabulary and data sources to improve the general perception of the model and its flexibility in real-life open-vocabulary recognition scenarios. These solutions assist to unlock extra practical capability of open-vocabulary recognition in the industry, including intelligent security systems, autonomous cars, robotic perception.

References

[1] Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE, 111*(3), 257-276.

[2] Jocher, G., Chaurasia, A., & Qiu, J. (2023). YOLO by Ultralytics. *Ultralytics*. https://github.com/ultralytics/ultralytics

[3] Ren, S., He, K., Girshick, R., & Sun, J. (2020). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1137-1149.

[4] Zaidi, S. S. A., Ansari, M. S., Aslam, A., Kanwal, N., Asghar, M., & Lee, B. (2022). A survey of modern deep learning based object detection models. *Digital Signal Processing, 126*, 103514.

[5] Zareian, A., Rosa, K. D., Hu, D. H., & Chang, S. F. (2021). Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14393-14402.

[6] Gu, X., Lin, T. Y., Kuo, R. J., & Cui, Y. (2022). Open-vocabulary object detection via vision-language pretrained bounding box supervision. In *European Conference on Computer Vision (ECCV)*, 521-537.

[7] Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 8748-8763.

[8] Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., & Ishan, M. (2022). Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision (ECCV)*, 350-368.

[9] Wu, J., Lu, T., Zhang, X., & Bai, Y. (2023). F-VLM: Open-vocabulary object detection upon frozen vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7083-7093.

DOI: 10.54254/2755-2721/2026.38494

[10] Yao, L., Han, J., Liang, X., Dan, Z., & Xu, C. (2022). DetCLIP: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems (NeurIPS), 35*, 9125-9138.
[11] Gupta, A., Dollar, P., & Girshick, R. (2020). LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5942-5951.
[12] Lin, C., Sun, P., Jiang, Y., Luo, P., Yuan, C., & Wang, J. (2023). Learning open-vocabulary object detection with multi-modal prompts. *IEEE Transactions on Multimedia, 25*, 8812-8824.
[13] Minderer, M., Zhang, A., Steiner, H., Li, Y., Weissenborn, D., & Unterthiner, T. (2022). Simple open-vocabulary object detection. *European Conference on Computer Vision (ECCV)*, 510-527.
[14] Liu, S., Li, Z., Zhang, C., & Yang, J. (2023). Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499.
[15] Wang, L., Zhang, Y., & Liu, X. (2024). Context-aware prompts for fine-grained open-vocabulary vision recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(2), 1102-1115.