Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ28957

Methods and Advances of Zero-Shot Object Detection

Jun Ou

Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University (BNBU), Zhuhai, China
t330026125@mail.uic.edu.cn

Abstract. Object detection is a relatively important task in the field of computer vision, and it has been widely applied in many popular fields in recent years. However, object detection still has many problems, such as time-consuming and labor-intensive dataset processing etc. Therefore, the zero-shot object detection method has been proposed. This paper classifies zero-shot object detection methods into two categories, which are classifier-based and instance-based methods, based on different problem-solving perspectives. It respectively introduces traditional semantic alignment and cross-modal alignment based on classifiers, as well as generative feature synthesis, pseudo-label mining, and knowledge distillation based on instances. Through specific cases, the actual research status, advantages, and disadvantages of three methods, namely cross-modal alignment, pseudo-labeling, and knowledge distillation, were analyzed in detail. This paper summarizes the problems and challenges that need to be urgently addressed in zero-shot object detection. An imbalance between positive and negative samples, significant semantic noise interference, and difficulty in visual-semantic alignment have restricted the further development of the research. It also introduces some solutions and future development directions.

Keywords: Zero-shot object detection, classifier-based, instance-based

1. Introduction

Object detection is a relatively important task in the field of computer vision. In recent years, this task has performed very well in some areas such as autonomous driving and video surveillance. This is attributed to the advancements in the field of deep learning and the continuous expansion of datasets used for training. However, the data used for training requires a large number of human resources to process and precisely label. The insufficiency of the dataset may lead to a reduction in the performance of the method. Furthermore, classifications that are not in the dataset cannot be correctly identified.

To address the above issues, a zero-shot object detection method was proposed, which has achieved remarkable results. As Ankan Bansal proposed, the Latent Assignment Based (LAB) method achieved a recall rate of 5.40% on Visual Genome, which was higher than the baseline (5.19%) [1]. Zang et al proposed a zero-shot object detection method for aerial images based on visually enhanced semantic vectors, achieving new optimal performance in the task of zero-shot

© 2025 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.TJ28957

object detection for aerial images [2]. At present, zero-shot object detection faces some challenges, mainly including noise, visible class bias, and computational power issues etc.

This review will categorize zero-shot object detection methods into two types from different perspectives of problem-solving, including classifier-based and instance-based. It will respectively introduce the focuses of the two methods and their respective approaches to achieving zero-shot functionality, and analyze the existing problems and future development directions.

2. Method

The key to classifier-based methods is to replace the classifier in the original method with one that has zero-shot learning ability. The existing methods can be divided into two subclasses based on the usage of different Spaces: traditional semantic alignment and cross-modal alignment.

2.1. Classifier-based method

The characteristic of traditional semantic alignment is to use the vision-semantic embedding framework to map the information in the visual space to the semantic space, achieving the knowledge transfer from known categories to unknown categories. This method has problems such as an ambiguous definition of background classes, a limited number of training classes that are difficult to cover the semantic space, and centrality. Therefore, the baseline zero-shot detection used by Ankan Bansal proposes a static background and latent allocation. These two methods solve background problems in different ways. Dense sampling can solve the problem of sparse semantic space by expanding the training categories [3].

To address the issue of centrality, cross-modal alignment was proposed. This type of method is no longer confined to the mapping from the visual space to the semantic space. It can achieve bidirectional association and integration between the visual and semantic spaces, and has a higher degree of utilization of information. Open vocabulary object detection also belongs to the method of cross-modal alignment. The basis for its implementation is the Visual-Language Model (VLM). VLM can handle information in both the visual and semantic spaces simultaneously. Through data training of large-scale image-text pairs, it maps the features of the visual and semantic spaces to the shared space, thereby establishing the association between the two types of features. Therefore, it can realize the understanding and generation of cross-modal [4].

However, existing VLMS such as CLIP are mainly used for aligning images with text and cannot meet the classification tasks of zero-shot object detection. Therefore, the DetCLIP model and the GLIP model have improved them from different directions. For instance, Yao et al believe that DetCLIP is an open vocabulary detection pre-training framework that adopts a dictionary-enhanced visual-concept parallelized expression. It integrates detection, grounding, and image-text pairs under a unified training objective and introduces a pseudo-label strategy to utilize network-scale image-text data. It also adopts a teacher-student distillation mechanism [5]. Li et al mentioned that GLIP unifies the object detection and phrase grounding tasks into one framework. It adopts deep cross-modal fusion and is pre-trained on 27 million data points, including manually labeled grounding data and web-crawled image-text pairs, possessing strong zero-shot and few-shot migration capabilities [6].

These two methods represent the improvement directions of the original structure from different directions. DetCLIP adopts the idea of region-text alignment, which enhances the model's contrastive learning ability and makes it more capable of distinguishing similar objects. GLIP

DOI: 10.54254/2755-2721/2026.TJ28957

unifies and fuses tasks, transforming the detection task into an alignment problem of images and text phrases, and further enhances the alignment capability by using deep cross-modal fusion.

2.2. Instance-based method

The core of the instance-based method is to address the issue of performance degradation caused by data loss. The key lies in generating labeled instances for unlabeled classes, enabling the model to conduct supervised learning to achieve the desired effect. However, the instances generated by this method are difficult to ensure the consistency between the synthesized and the real invisible classes. At the same time, it is hard to locate the positions of the invisible classes in the image. These two problems need to be solved by the instance-based method. The main solution of generative feature synthesis is to generate synthetic visual features for invisible classes through semantic information. For example, Pengkai Zhu introduced a visual consistency checker and imposed triple constraints on the generated features to ensure that the confidence score of the generated features was consistent with the real features [7]. This method solves the problem of a lack of data to a certain extent. However, such methods only update the classifier module and ignore the positioning module in the detection model, which has a certain impact on the positioning ability of the model.

The solution to pseudo-label mining is to generate pseudo-labels for unlabeled images, enabling the model to utilize this data. The VL-PLM method proposed by Shiyu Zhao utilizes the rich semantic information in the vision-language model to locate and classify targets in unlabeled images, thereby generating pseudo-labels for target detection. The method uses these models to improve the existing object detection frameworks that utilize unlabeled data [8]. Meanwhile, VL-PLM fuses the semantic score of the vision-language model with the localization score of the regional proposal network, optimizes the regional boundary, and improves the localization accuracy of pseudo-labels.

Knowledge distillation is another approach to solving problems. It distills the knowledge of the visual-language model and transfers it to the detector. The former is called the teacher model, and the latter is called the student model. The ViLD method proposed by Xiuye Gu uses the teacher model to encode the category text and image regions, and then trains a detector so that the region embeddings of its detection box are consistent with the text and image embeddings inferred by the teacher model. And the method replaces the class-dependent location module in the detector with a class-independent module to generalize to unseen classes [9].

The above two types of zero-shot object detection are different perspectives on solving problems. Classifier-based methods typically attempt to learn a mapping function from the visual feature space to the semantic space. The fundamental purpose is to utilize the visual and semantic information of known categories to train a classifier capable of understanding and recognizing the semantics of unknown categories. The instance-based methods involve creating or finding representative visual instances for each unseen category, transforming zero-shot object detection into a relatively mature, supervised few-shot or traditional detection problem.

3. Case analysis

3.1. GLIP

GLIP is a unified pre-training method for open vocabulary object detection. Its core is to unify object detection as a phrase localization task. It means treating each category in the detection task as a text phrase and achieving open vocabulary detection through alignment between images and

DOI: 10.54254/2755-2721/2026.TJ28957

language. During detection, GLIP simultaneously inputs image and text prompts, and then outputs the regional image features that match the text phrases. The similarity between the regional image features and the text features is calculated through the dot product, replacing the traditional classification head. At the same time, a cross-attention module is inserted between the image encoder and the language encoder, enabling each modality to focus on the information of the other modality. It realizes the visual understanding guided by language.

Under the zero-shot setting, GLIP achieved an average precision (AP) of 49.8 on the COCO validation set without any pre-training of COCO images, while the traditional Faster R-CNN was 40.2 AP. It is nearly 10 points higher [6]. On the LVIS dataset containing thousands of rare targets, GLIP-Large still had 26.9 AP, outperforming Supervised-RFS, which was fully supervised and trained on LVIS [6]. This detection method has high versatility. It only requires modifying the text prompt and can be applied to any category. Moreover, it has high deployment efficiency and does not require retraining the model, supporting prompt fine-tuning. However, this method is highly dependent on language models. The design of prompts can affect performance, and manual optimization of prompts is required to achieve high performance. In addition, this method requires large-scale pre-training of image-text pairs. It has a high training cost and is difficult to capture small targets. GLIP was the first to unify the detection and location tasks, promoting the paradigm "image-text large model + detection". It lays the foundation for the subsequent OVD method and becomes a significant work in the field of open vocabulary detection.

3.2. VL-PLM

VL-PLM is a method for zero-shot object detection using pseudo-labels. Its core lies in using the semantic capabilities of large-scale vision-language models to generate pseudo-labels on unlabeled images. First, the VL-PLM trains the class-independent proposal generator, using only the real boxes of the base class (such as Class COCO 48) for training, without the need for new class (such as Class COCO 17) annotations. Then, it uses the text encoder of CLIP to encode the new class names into text embeddings. Finally, the final detector is trained with real annotations of the base class and pseudo-labels of the new class.

For the zero-shot dataset of COCO 2017, a standard Mask R-CNN was trained with pseudo-labels generated by VL-PLM. On the test set, the average accuracy (AP50) of the new category reached 34.4, while the previous best method, ViLD, was only 27.6. It's equivalent to an increase of 6.8 points. Meanwhile, the accuracy of the old base classes was hardly compromised and remained at around 60.2 [8]. This method has strong universality. The unified framework is adapted to OVD and SSOD, significantly enhancing the performance of new class detection. The performance of the base class remains almost unchanged, and it has strong generalization ability across datasets. It does not require new class training data and can also be orthogonally combined with other SSOD methods to further improve performance. However, the pseudo-label method relies on the local localization capability of CLIP, and problems such as "partial boxes", "redundant boxes", and "missed detections" still occur. Meanwhile, the generation of pseudo-labels takes a considerable amount of time. Although there is a Fast version, it is still slower than traditional methods. It is sensitive to the modeling method of background categories and requires manual design of category spaces, which has potential for improvement. VL-PLM inspired subsequent work, integrating the V&L model more deeply with detection tasks. The pseudo-label generation paradigm was borrowed by subsequent semi-supervised detection work to improve the quality of pseudo-labels.

DOI: 10.54254/2755-2721/2026.TJ28957

3.3. Knowledge distillation

Knowledge distillation is a unique and innovative method. ViLD breaks down open vocabulary detection into "general target localization" and "open vocabulary classification". First, it trains a class-agnostic RPN using the base class to generate candidate boxes, and then distills the knowledge of CLIP/ALIGN in real time into a new detector through Vilid-Text and Vilid-Image. This new detector is the student model, while the pre-trained open-vocabulary image-text alignment model, such as CLIP or ALIGN, is the teacher model, which is responsible for providing both text embedding and image embedding "knowledge". The distillation samples of ViLD-image may contain unlabeled novel objects, so the embedding space learned by the student model is naturally compatible with unseen categories, and the entire process does not require any training data of novel categories.

On the LVIS dataset, ViLD was trained with only 866 common categories and achieved a mask APr of 16.1 on 337 rare categories that had never been seen before, which was 3.8 points higher than the fully supervised baseline under the same conditions [9]. If the teacher model is changed from CLIP to a stronger one, ALIGN, the novel AP jumps directly to 26.3, narrowing the gap with the LVIS Challenge champion of that year (30.0) to 3.7 [9]. This method does not require novel category training data, which is a true zero-shot detection, and has verified its cross-domain migration ability on multiple datasets. It was the first to introduce the knowledge distillation of CLIP style into object detection. The method proposed the ViLD framework, solved the core problem of "difficult data expansion" in open vocabulary detection, and inspired subsequent research. However, in this method, the performance of the student model depends on that of the teacher model. It is sensitive to regions with extreme aspect ratios and has limited ability to recognize fine-grained categories. It is also prone to confusion regarding visually similar categories, still lagging behind fully supervised methods.

4. Challenge

At present, zero-shot object detection has made certain progress and development, but there are still many challenges that limit its further development. For instance, there is an imbalance between positive and negative samples, significant semantic noise interference, and difficulties in visual-semantic alignment, etc. Whether it is the instance-based detection method or the classifier-based detection framework, they are all affected by the above problems to varying degrees.

Visual-semantic alignment is currently the focus of zero-shot object detection. Classifier-based methods rely on semantic embeddings, but these embeddings are not always consistent with visual features. Even if instance-based methods generate visual features, they still need to be guided by semantic labels. If the semantic embedding is inaccurate, the generated samples will also deviate from the true distribution. Due to the lack of data on new categories, the model has difficulty learning the accurate mapping between visual features and text features, resulting in possible misclassification of objects [2]. At the same time, the problem of high noise in the semantic space also lies in the use of semantic embeddings, which come from unsupervised texts with high noise. This is an inherent issue and has nothing to do with whether samples are generated or not.

In addition, the difficulty in detecting small targets is also a major challenge. The small targets existing in the training set are hard to detect, and there is no category or training support. For instance-based methods, the generator has difficulty generating high-quality small target samples, which often leads to errors in tasks related to small targets. Tan pointed out that the imbalance of positive and negative samples is a problem that exists in the zero-shot object detection task. Due to

DOI: 10.54254/2755-2721/2026.TJ28957

the lack of labeled data for unseen categories, it is difficult to distinguish the background from the unseen categories. The detector will produce fewer positive samples and more negative samples [1]. This has a significant impact on classifiers. Although instance-based methods can partially alleviate this by generating new category samples, the limited number of generated samples still makes it difficult to strike a balance.

5. Future direction of development

For the noise problem, attempts can be made to mitigate the impact of noise on the classifier. For example, a mixed attribute space of manually defined and automatically learned by the network can be used. Metric or contrastive learning can be carried out in the shared space to mitigate the domain offset caused by text noise [1]. And attempts can also be made to optimize the method of cross-modal alignment. Through multi-stage cascading, each stage uses semantic information to re-score. And they need to gradually eliminate non-target noise boxes to reduce the interference of noise on the final classifier. Meanwhile, how to obtain purer and more visually aligned semantic representations and reduce noise from the source are also the current research focuses.

At present, the common problems of performance degradation of old classes and catastrophic forgetting have restricted further development. Since the open-world approach inherently has the ability to cover a wider range of categories compared to the zero-shot paradigm [2], ZSD will be more embedded in open-world incremental learning systems in the future to address the problems of catastrophic forgetting and continuous scaling. For the performance degradation of old classes, it is necessary to design a unified framework in the future that ensures old classes are not forgotten and new classes are detectable. One can attempt to design a loss function (such as polarity loss) that considers both visible and invisible classes, and adopt a cascaded region proposal network. This approach can maintain the performance of old classes while improving the recall rate of new classes and alleviating the performance degradation of old classes [1]. Alternatively, a cascaded structure can be used to gradually refine candidate boxes and reduce the false suppression of old classes.

6. Conclusion

This paper provides a detailed introduction to zero-shot object detection. Firstly, this paper introduces the background of zero-shot object detection. This article also summarizes the definition of zero-shot object detection. Subsequently, this paper classifies the existing methods into two categories based on the different perspectives from which zero-shot object detection solves problems, namely classifier-based methods and instance-based methods. The classifier-based method mainly introduces traditional semantic alignment and cross-modal alignment, and the difference between them lies in the usage of different Spaces. The instance-based methods are introduced from three different perspectives: generative feature synthesis, pseudo-label mining, and knowledge distillation. In addition, this paper presents three specific examples to analyze the principles, advantages, and disadvantages of these methods. Open vocabulary has high universality, but the training cost is relatively high. Pseudo-labels have strong generalization ability across datasets, rely on local localization of Clips, and take a long time. Knowledge distillation is true zero-shot object detection, but its ability is limited. Finally, this paper discusses the main problems faced by zero-shot object detection, such as the imbalance of positive and negative samples, large semantic noise interference, and difficulty in visual-semantic alignment. It introduces the corresponding solutions and future development directions.

DOI: 10.54254/2755-2721/2026.TJ28957

References

[1] Tan, C., Xu, X., & Shen, F. (2021). A Survey of zero shot detection: Methods and applications. Cognitive Robotics, 1, 159–167.
[2] Cao, W., Yao, X., Xu, Z., Liu, Y., Pan, Y., & Ming, Z. (2025). A Survey of Zero-Shot Object Detection. Big Data Mining and Analytics, 8(3), 726–750.
[3] Bansal, A., Sikka, K., Sharma, G., Rama Chellappa, & Ajay Divakaran. (2018). Zero-Shot Object Detection. Lecture Notes in Computer Science, 397–414.
[4] Gupta, D., Anantharaman, A., Nehal Mamgain, Sowmya Kamath S, Balasubramanian, V. N., & C.V. Jawahar. (2020). A Multi-Space Approach to Zero-Shot Object Detection. 1198–1206.
[5] Yao, L., Han, J., Liang, X., Xu, D., Zhang, W., Li, Z., & Xu, H. (2023). DetCLIPv2: Scalable Open-Vocabulary Object Detection Pre-training via Word-Region Alignment. ArXiv (Cornell University).
[6] Liunian Harold Li, Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Liu, Y., Zhang, L., Hwang, J., Chang, K.-W., & Gao, J. (2022). Grounded Language-Image Pre-training. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[7] Zhu, P., Wang, H., & Saligrama, V. (2020). Don't Even Look Once: Synthesizing Features for Zero-Shot Detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11690–11699.
[8] Zhao, S., Zhang, Z., Schulter, S., Zhao, L., Kumar, B. G. V., Stathopoulos, A., Manmohan Chandraker, & Metaxas, D. N. (2022). Exploiting Unlabeled Data with Vision and Language Models for Object Detection. Lecture Notes in Computer Science, 159–175.
[9] Gu, X., Lin, T.-Y., Kuo, W., & Cui, Y. (2021, April 1). Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. NASA ADS.