A Survey on Vision-Language Models and Object Detection

Yiliang Chen, Qingzheng Wang, Shubing Yan
Information Engineering, North China University of Water Resources and Electric Power, China

ABSTRACT: *This paper reviews the latest research advances in vision-language large models and the field of object detection. In recent years, Vision-Language Models (VLMs) have made significant progress at the intersection of computer vision and natural language processing, demonstrating great potential in object detection tasks. Traditional object detection methods rely on large amounts of annotated data and predefined categories, whereas vision-language models enhance detection capabilities by leveraging multimodal contrastive learning and cross-modal alignment, especially excelling in open-vocabulary detection, zero-shot detection, and few-shot learning tasks. This paper first introduces the fundamental concepts and development history of vision-language large models and object detection, followed by a detailed discussion of their applications in object detection.*

KEYWORDS *-Vision-Language Large Models, Object Detection, Open-Vocabulary Detection*

I. INTRODUCTION

In recent years, with the rapid development of deep learning and artificial intelligence technologies, vision-language large models and object detection techniques have made significant progress, becoming research hotspots in the fields of computer vision and natural language processing. Vision-language large models, by integrating computer vision and natural language processing techniques, can understand and generate text descriptions related to images, and are widely applied in tasks such as image captioning, visual question answering, and cross-modal retrieval. As one of the core tasks in computer vision, object detection aims to identify and locate specific objects in images, providing rich visual information for vision-language large models. The combination of these two technologies not only drives the advancement of multimodal artificial intelligence but also offers strong technical support for real-world applications.

The integration of vision-language large models with object detection technology has opened up new possibilities for the advancement of multimodal artificial intelligence. On one hand, vision-language large models can leverage object detection techniques to better understand image content, thereby generating more accurate textual descriptions. For example, in image captioning tasks, object detection can help models identify key objects in an image and their relationships, leading to richer and more precise textual descriptions [1]. On the other hand, object detection can benefit from vision-language large models by utilizing textual information to enhance detection performance. For instance, in open-vocabulary object detection (OVOD) tasks, vision-language large models can expand detection categories through textual descriptions, enabling the detection of previously unseen classes [2].

Despite the significant progress in integrating vision-language large models with object detection technology, several challenges and issues remain. First, the construction of large-scale multimodal datasets is a critical issue. Although existing vision-language datasets (such as COCO and Visual Genome) are relatively large, they still have limitations in terms of diversity and coverage [3, 4]. Second, computational efficiency and real-time performance are key concerns. Vision-language large models and object detection techniques typically require substantial computational resources, making it difficult to meet real-time requirements in practical applications [5].

Additionally, the generalization capability and robustness of these models remain significant challenges. Current vision-language large models and object detection techniques exhibit limitations when handling complex scenes and unseen categories, necessitating further improvements [6].

In conclusion, the integration of vision-language large models with object detection technology has opened up new possibilities for the development of multimodal artificial intelligence, promoting the fusion and advancement of computer vision and natural language processing. This paper aims to review the mutually beneficial relationship between vision-language large models and object detection, exploring the latest research progress in their integration.

II. VISION-LANGUAGE MODELS

Early vision-language models were primarily based on CNN and RNN, which encoded images and text into feature vectors separately and then performed simple feature fusion. For example, the "Show and Tell" model proposed by Kiros et al. used a CNN to extract image features and an RNN to generate corresponding text descriptions [7]. Another early representative work is the "Neural Image Captioning" model proposed by Vinyals et al., which also adopted a CNN-RNN architecture, feeding image features into an RNN to generate text descriptions [8].

With the introduction of the Transformer architecture, research on vision-language models entered a new stage. The Transformer, through its self-attention mechanism, enabled parallel processing of sequential data, significantly enhancing the model's expressive power and training efficiency. For example, the Transformer model proposed by Vaswani et al. achieved groundbreaking progress in machine translation tasks [9]. Vision-language models based on the Transformer, such as CLIP[10] (in Figure 1) and OFA[11], learned rich visual and linguistic representations by pretraining on large-scale image-text pair datasets, significantly improving the performance of multimodal tasks.The ALIGN[12] model achieves zero-shot learning for unseen categories and tasks by pretraining on a large-scale image-text pair dataset .

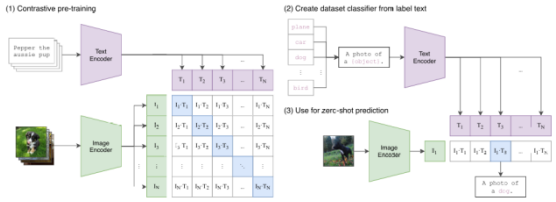

Fig.1 CLIP model architecture

Self-supervised learning and contrastive learning are another important direction in the recent research of vision-language large models. Self-supervised learning designs pretraining tasks to utilize unlabeled data for model training, thereby reducing the reliance on labeled data. For example, the SimCLR model uses contrastive learning to achieve self-supervised learning of image features, significantly improving performance in image classification and object detection tasks [13].In vision-language large models, contrastive learning is widely used for multimodal feature alignment and representation learning. For instance, the ALBEF model uses contrastive learning to achieve fine-grained alignment of image and text features, resulting in better performance in visual question answering and image captioning tasks [14]. This type of research shows that contrastive learning is an effective strategy for improving the performance of vision-language large models.

Despite the significant performance improvements in vision-language large models, their computational complexity and resource requirements have also increased significantly, making it difficult to meet real-time requirements in practical applications. As a result, model compression and knowledge distillation have become important directions in vision-language large model research. For example, the DistilBERT model uses knowledge distillation to compress large-scale pre-trained language models into smaller models, significantly reducing computational complexity while maintaining performance [15].

In vision-language large models, knowledge distillation is widely applied for model compression and acceleration. For instance, the TinyVL model uses knowledge distillation to compress large-scale vision-language models into smaller models, enabling efficient deployment on mobile devices and embedded systems [16]. This type of research demonstrates that model compression and knowledge distillation are important techniques for enhancing the practical

application value of vision-language large models.

III. OBJECT DETECTION

With the rise of deep learning technology, object detection has further developed into two main categories: single-stage detectors and two-stage detectors. Single-stage detectors achieve faster detection speeds by directly performing dense sampling and classification on the image. For example, the YOLO (You Only Look Once) method proposed by Redmon et al. treats object detection as a regression problem, completing the task with a single forward pass, which significantly improves detection speed [17]. The YOLO series methods (such as YOLOv3 and YOLOv4) further enhance detection accuracy and speed by introducing multi-scale prediction and Feature Pyramid Networks (FPN) [18, 19].Another representative single-stage detector is the SSD (Single Shot MultiBox Detector) method proposed by Liu et al., which achieves effective detection of objects at different scales by making predictions across multiple feature maps [20]. The SSD method combines multi-scale features with a default box mechanism, significantly improving detection accuracy while maintaining high detection speed.

Two-stage detectors achieve higher detection accuracy by first generating region proposals and then performing fine classification and regression. The R-CNN (Region-based Convolutional Neural Networks) series methods, proposed by Girshick et al., are pioneering works in deep learning-based object detection [21]. R-CNN generates candidate regions using Selective Search and extracts region features using CNNs, followed by classification and bounding box regression with an SVM classifier. To improve detection efficiency, Girshick et al. proposed the Fast R-CNN method, which introduced the RoI pooling layer to enable shared computation of candidate region features, significantly reducing computational cost [22]. Later, Ren et al. proposed the Faster R-CNN method, which introduced the Region Proposal Network (RPN) to further unify candidate region generation and object detection into an end-to-end framework, significantly improving both detection speed and accuracy [23].

IV. OPEN VOCABULARY OBJECT DETECTION

Open Vocabulary Object Detection (OVOD) is an important research direction in the field of object detection, aiming to address the dependency of traditional object detection methods on fixed category labeled data. Traditional object detection methods can typically only detect categories that appear in the training set, while open vocabulary object detection requires the model to detect unseen categories. In recent years, with the development of large-scale vision-language models such as CLIP, significant progress has been made in open vocabulary object detection. The CLIP model, by pretraining on a large-scale image-text paired dataset, has learned powerful cross-modal alignment capabilities, allowing it to map images and texts to the same semantic space, thereby enabling open vocabulary object detection.

Open Vocabulary Object Detection (OVOD) leverages the concept of open vocabulary learning, using image-text knowledge to train on known category data, thereby enabling detection of unseen categories. This approach acquires rich knowledge by utilizing a large amount of additional data to cover more object detection categories, and transfers this knowledge to a general object detection framework for further training. This allows a closed-set object detector to be extended into an open-vocabulary object detector, enabling it to recognize and detect new categories that were not seen during training.

4.1 REGION-TEXT PRETRAINING

Region-Text Pretraining is an important approach in Open-Vocabulary Object Detection (OVOD), aiming to align image regions with textual descriptions by pretraining on large-scale region-text paired data, thereby enabling the detection of unseen categories. The core idea is to leverage vision-language models (such as CLIP) to map image region features and text features into the same semantic space. Through contrastive learning or other alignment mechanisms, the model can detect categories not present in the training data based on textual descriptions. Specifically, Region-Text Pretraining typically involves the following steps: first, extracting image region features using object detection frameworks (such as Faster R-CNN or DETR); second, extracting text features using pretrained language models (such as BERT or CLIP text encoders); then, aligning

region features and text features through contrastive learning or cross-modal attention mechanisms; finally, during inference, detecting unseen categories by computing the similarity between image region features and text features. This approach not only significantly enhances the model's generalization ability for open-vocabulary scenarios but also avoids the high cost of retraining the model.

OVR-CNN[24] utilizes large-scale region-text paired data (such as the Visual Genome dataset) for pretraining, mapping image region features and textual descriptions into the same semantic space. Through this approach, the model learns the semantic relationships between image regions and text. This method enables the extension of detectable categories without the need to retrain the model.To better capture the correspondence between regions and text, MEDet[25] jointly trains an object detector using mini-batches of data from both detection datasets and image caption datasets. During training, the caption text is parsed to extract word-level text, which may contain information about new categories.RegionCLIP[26] uses pseudo-label text alignment to obtain pseudo-region-text pairs, which are then input into the model to pretrain the image encoder to learn region information. The visual encoder is fine-tuned using a manually annotated dataset to adapt to different detection tasks.

4.2 KNOWLEDGE DISTILLATION

In open vocabulary object detection tasks, knowledge distillation can help enhance the capabilities of the student model by learning from the knowledge of a complex teacher model, especially when computational resources are limited. Specifically, the teacher model is typically a large-scale vision-language model that has been pre-trained on vast amounts of image and text data, enabling it to effectively recognize various object categories and understand cross-modal information. Through knowledge distillation, the student model can learn how to handle the relationship between images and text from the teacher model, thereby improving its performance in open vocabulary object detection tasks.

The ViLD[27] model uses knowledge distillation to transfer image and text knowledge from a pre-trained open-vocabulary image classification model to a two-stage detector, addressing the problem of limited training data in open-vocabulary object detection (OVD) tasks. The ViLD model employs a vision-language model (VLM) image encoder to compute image embeddings for cropped regions and a text encoder to obtain text embeddings for categories. These text embeddings are then used as inputs to the region classifier. This approach allows ViLD to better handle the challenges of category expansion and cross-modal learning in open-vocabulary object detection tasks.

4.3 TRANSFER LEARNING

Transfer learning typically involves fine-tuning a large pre-trained model or extracting visual features for downstream tasks. The F-VLM[28] model adopts a transfer learning-based approach. The main feature of this method is the direct use of a pre-trained vision-language model (VLM) image encoder to train the detection head, which is matched with the text features generated by the CLIP text encoder. A Region Proposal Network (RPN) is used to generate candidate regions for feature extraction. The extracted region embeddings are then compared with the text embeddings of the CLIP text encoder.

4.4 PROMPT LEARNING

Prompt learning is a method that guides th e model's learning using text prompts. In the field o f image recognition, prompts can be image descript ions, labels, or classification information. By desig ning these prompts effectively, they can adjust and optimize the base model, enabling it to adapt to var ious downstream tasks.

To obtain the text embeddings of class na mes, prompts are input into the pretrained VLM tex t encoder to generate them, and these embeddings a re used to supervise the training of the region classi fier for object detection. The PromptDet[29] model incorporates a series of learnable vectors into the t ext input. These vectors do not correspond to any a ctual words, but instead serve as virtual tokens to h elp better align the text embedding space with the v isual representations for object detection.

V. CONCLUSION

The integration of vision-language models with object detection has brought revolutionary advancements to the field of computer vision, especially in the OVOD task. By combining the

powerful semantic understanding capabilities of vision-language models (such as CLIP and ALIGN) with object detection frameworks, researchers have successfully enabled the detection of unseen categories, significantly enhancing the model's generalization ability. Core methods include region-text alignment, knowledge distillation, prompt learning, and pseudo-label generation, among others. These techniques, through multimodal fusion and semantic alignment, enable the model to flexibly extend detection categories using text descriptions. Representative works such as RegionCLIP, ViLD, and FVLM have driven improvements in open-vocabulary detection performance. In the future, as multimodal pretraining technologies continue to develop, the integration of vision-language models with object detection will become even more seamless, potentially achieving greater breakthroughs in zero-shot learning, few-shot learning, and complex scene understanding.

REFERENCES

[1] Anderson, P., He, X., Buehler, C., et al., "Bottom-up and top-down attention for image captioning and visual question answering," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2018, pp. 6077-6086.

[2] Gu, X., Lin, T. Y., Kuo, W., and Cui, Y., "Open-vocabulary object detection via vision and language knowledge distillation," arXiv preprint arXiv:2104.13921, 2021.

[3] Lin, T. Y., Maire, M., Belongie, S., et al., "Microsoft COCO: Common objects in context," Proc. European Conf. on Computer Vision, 2014, pp. 740-755.

[4] Krishna, R., Zhu, Y., Groth, O., et al., "Visual Genome: Connecting language and vision using crowdsourced dense image annotations," Int. J. Computer Vision, vol. 123, no. 1, pp. 32-73, 2017.

[5] Howard, A. G., Zhu, M., Chen, B., et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[6] Hendrycks, D., and Dietterich, T., "Benchmarking neural network robustness to common corruptions and perturbations," arXiv preprint arXiv:1903.12261, 2019.

[7] Kiros, R., Salakhutdinov, R., and Zemel, R., "Unifying visual-semantic embeddings with multimodal neural language models," arXiv preprint arXiv:1411.2539, 2014.

[8] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D., "Show and tell: A neural image caption generator," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2015, pp. 3156-3164.

[9] Vaswani, A., Shazeer, N., Parmar, N., et al., "Attention is all you need," Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[10] Radford, A., Kim, J. W., Hallacy, C., et al., "Learning transferable visual models from natural language supervision," Proc. Int. Conf. on Machine Learning, 2021, pp. 8748-8763.

[11] Wang, P., Yang, A., Men, R., et al., "OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," Proc. Int. Conf. on Machine Learning, 2022, pp. 23318-23340.

[12] Jia, C., Yang, Y., Xia, Y., et al., "Scaling up visual and vision-language representation learning with noisy text supervision," Proc. Int. Conf. on Machine Learning, 2021, pp. 4904-4916.

[13] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G., "A simple framework for contrastive learning of visual representations," Proc. Int. Conf. on Machine Learning, 2020, pp. 1597-1607.

[14] Li, J., Selvaraju, R., Gotmare, A., et al., "Align before fuse: Vision and language representation learning with momentum distillation," Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 9694-9705.

[15] Sanh, V., Debut, L., Chaumond, J., and Wolf, T., "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

[16] Wang, W., Bao, H., Dong, L., et al., "TinyVL: A tiny vision-language model for efficient multimodal learning," arXiv preprint arXiv:2109.10380, 2021.

[17] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A., "You only look once: Unified, real-time object detection," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2016, pp. 779-788.

[18] Redmon, J., and Farhadi, A., "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[19] Bochkovskiy, A., Wang, C. Y., and Liao, H. Y. M., "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[20] Liu, W., Anguelov, D., Erhan, D., et al., "SSD: Single shot multibox detector," Proc. European Conf. on Computer Vision, 2016, pp. 21-37.

[21] Girshick, R., Donahue, J., Darrell, T., and Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2014, pp. 580-587.

[22] Girshick, R., "Fast R-CNN," Proc. IEEE Int. Conf. on Computer Vision, 2015, pp. 1440-1448.

[23] Ren, S., He, K., Girshick, R., and Sun, J., "Faster R-CNN: Towards real-time object detection with region proposal networks," Adv. Neural Inf. Process. Syst., vol. 28, 2015, pp. 91-99.

[24] Zareian, A., Dela Rosa, K., Hu, D. H., and Chang, S.-F., "Open-Vocabulary Object Detection Using Captions," Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2021.

[25] Chen, P., Sheng, K., Zhang, M., et al., "Open Vocabulary Object Detection with Proposal Mining and Prediction Equalization," arXiv preprint arXiv:2206.11134, 2022.

[26] Zhong, Y., Yang, J., Li, C., Zhang, P., Gao, J., and Zhang, L., "RegionCLIP: Region-based Language-Image Pretraining," Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Piscataway, NJ, USA: IEEE, 2022, pp. 16793-16803.

[27] Gu, X., Lin, T.-Y., Kuo, W., and Cui, Y., "Open-vocabulary object detection via vision and language knowledge distillation," Proc. IEEE/CVF Int. Conf. on Computer Vision, Piscataway, NJ, USA: IEEE, 2021, pp. 14379–14388.

[28] Kuo, W., Cui, Y., Gu, X., Lin, T.-Y., and Zhang, Z., "FVLM: Open-vocabulary object detection upon frozen vision and language models," Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Piscataway, NJ, USA: IEEE, 2023, pp. 11241–11250..

[29] Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., and Ma, L., "PromptDet: Towards open-vocabulary detection using uncurated images," Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Piscataway, NJ, USA: IEEE, 2023, pp. 1–10.