Multiple Prompt Fusion for Zero-Shot Lesion Detection Using Vision-Language Models

Miaotian Guo[1], Huahui Yi[2], Ziyuan Qin[2], Haiying Wang[1], Aidong Men[1], and Qicheng Lao[1,3](✉)

[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
qicheng.lao@bupt.edu.cn

[2] West China Biomedical Big Data Center, West China Hospital, Sichuan University, Sichuan, China

[3] Shanghai Artificial Intelligence Laboratory, Shanghai, China

Abstract. The success of large-scale pre-trained vision-language models (VLM) has provided a promising direction of transferring natural image representations to the medical domain by providing a well-designed prompt with medical expert-level knowledge. However, one prompt has difficulty in describing the medical lesions thoroughly enough and containing all the attributes. Besides, the models pre-trained with natural images fail to detect lesions precisely. To solve this problem, fusing multiple prompts is vital to assist the VLM in learning a more comprehensive alignment between textual and visual modalities. In this paper, we propose an ensemble guided fusion approach to leverage multiple statements when tackling the phrase grounding task for zero-shot lesion detection. Extensive experiments are conducted on three public medical image datasets across different modalities and the detection accuracy improvement demonstrates the superiority of our method.

Keywords: Vision-language models · Lesion detection · Multiple prompts · Prompt fusion · Ensemble learning

1 Introduction

Medical lesion detection plays an important role in assisting doctors with the interpretation of medical images for disease diagnosing, cancer staging, etc., which can improve efficiency and reduce human errors [9,19]. Current object detection approaches are mainly based on supervised learning with abundant well-paired image-level annotations, which heavily rely on expert-level knowledge. As such, these supervised approaches may not be suitable for medical lesion detection due to the laborious labeling.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43904-9_28.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 283–292, 2023.
https://doi.org/10.1007/978-3-031-43904-9_28

Recently, large-scale pre-trained vision-language models (VLMs), by learning the visual concepts in the images through the weak labels from text, have prevailed in natural object detection or visual grounding and shown extraordinary performance. These models, such as GLIP [11], X-VLM [10], and VinVL [24], can perform well in detection tasks without supervised annotations. Therefore, substituting conventional object detection with VLMs is possible and necessary. The VLMs are first pre-trained to learn universal representations via large-scale unlabelled data and can be effectively transferred to downstream tasks. For example, a recent study [15] has demonstrated that the pre-trained VLMs can be used for zero-shot medical lesion detection with the help of well-designed prompts.

However, current existing VLMs are mostly based on a single prompt to establish textual and visual alignment. This prompt needs refining to cover all the features of the target as much as possible. Apparently, even a well-designed prompt is not always able to combine all expressive attributes into one sentence without semantic and syntactic ambiguity, e.g., the prompt design for melanoma detection should include numerous kinds of information describing attributes complementing each other, such as shape, color, size, etc [8]. In addition, each keyword in a single lengthy prompt cannot take effect equally as we expect, where the essential information can be ignored. This problem motivates us to study alternative approaches with multiple prompt fusion.

In this work, instead of striving to design a single satisfying prompt, we aim to take advantage of pre-trained VLMs in a more flexible way with the form of multiple prompts, where each prompt can elicit respective knowledge from the model which can then be fused for better lesion detection performance. To achieve this, we propose an ensemble guided fusion approach derived from clustering ensemble learning [3], where we design a step-wise clustering mechanism to gradually screen out the implausible intermediate candidates during the grounding process, and an integration module to obtain the final results by uniting the mutually independent candidates from each prompt. In addition, we also examine the language syntax based prompt fusion approach as a comparison, and explore several fusion strategies by first grouping the prompts either with described attributes or categories and then repeating the fusion process.

We evaluate the proposed approach on a broad range of public medical datasets across different modalities including photography images for skin lesion detection ISIC 2016 [2], endoscopy images for polyp detection CVC-300 [21], and cytology images for blood cell detection BCCD. The proposed approach exhibits extraordinary superiority compared to those with single prompt and other common ensemble learning based methods for zero-shot medical lesion detection. Considering the practical need of lesion detection, we further provide significantly improved fine-tuning results with a few labeled examples.

2 Related Work

Object Detection and Vision-Language Models. In the vision-language field, phrase grounding can be regarded as another solution for object detection apart from conventional R-CNNs [5,6,18]. Recently, vision-language models have achieved exciting performance in the zero-shot and few-shot visual recognition [4,16]. GLIP [11] unifies phrase grounding and object detection tasks, demonstrating outstanding transfer capability. In addition, ViLD [7] is proposed for open-vocabulary object detection taking advantage of the rich knowledge learned from CLIP [4] and text input.

Ensemble Learning. As pointed out by a review [3], ensemble learning methods achieve better performance by producing predictions based on extracted features and fusing via various voting mechanisms. For example, a selective ensemble of classifier chains [13] is proposed to reduce the computational cost and the storage cost arose in multi-label learning [12] by decreasing the ensemble size. UNDEED [23], a semi-supervised classification method, is presented to increase the classifier accuracy on labeled data and diversity on unlabeled data simultaneously. And a hybrid semi-supervised clustering ensemble algorithm [22] is also proposed to generate basic clustering partitions with prior knowledge.

3 Method

In this section, we first briefly introduce the vision-language model for unifying object detection as phrase grounding, e.g., GLIP [11] (Sect. 3.1). Then we present a simple language syntax based prompt fusion approach in Sect. 3.2. Finally, the proposed ensemble-guided fusion approach and several fusion strategies are detailed in Sect. 3.3 to improve the zero-shot lesion detection.

3.1 Preliminaries

Phrase grounding is the task of identifying the fine-grained correspondence between phrases in a sentence and objects in an image. The GLIP model takes as input an image I and a text prompt p that describes all the M candidate categories for the target objects. Both inputs will go through specific encoders Enc_{I} and Enc_{T} to obtain unaligned representations. Then, GLIP uses a grounding module to align image boxes with corresponding phrases in the text prompt. The whole process can be formulated as follows:

$$O = Enc_{\mathrm{I}}(I),\ P = Enc_{\mathrm{T}}(p),\ S_{\mathrm{ground}} = OP^{\top},\ L_{\mathrm{cls}} = Loss(S_{\mathrm{ground}};\ T), \tag{1}$$

where $O \in \mathbb{R}^{N\times d}$, $P \in \mathbb{R}^{M\times d}$ denote the image and text features respectively for N candidate region proposals and M target objects, $S_{\mathrm{ground}} \in \mathbb{R}^{N\times M}$ represents the cross-modal alignment scores, and $T \in \{0,\ 1\}^{N\times M}$ is the target matrix.

3.2 Language Syntax Based Prompt Fusion

As mentioned above, it is difficult for a single prompt input structure such as GLIP to cover all necessary descriptions even through careful designation of the prompt. Therefore, we propose to use multiple prompts instead of a single

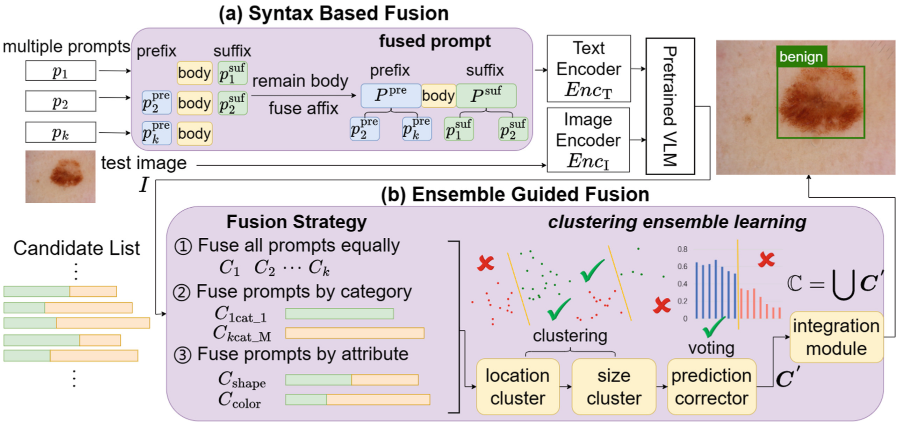

Fig. 1. Overview of the proposed approach: (a) Syntax Based Fusion, which fuses multiple prompts at the prompt level; (b) Ensemble Guided Fusion, which includes step-wise clustering mechanisms followed by voting and an integration module.

prompt for thorough and improved grounding. However, it is challenging to combine the grounding results from multiple prompts since manual integration is subjective, ineffective, and lacks uniform standards. Here, we take the first step to fuse the multiple prompts at the prompt level. We achieve this by extracting and fusing the prefixes and suffixes of each prompt based on language conventions and grammar rules. As shown in Fig. 1 (a), given serials of multiple prompts $p_1, p_2, \ldots, p_k$, the final fused prompt P_{fuse} from k single prompts is given by:

$$\begin{aligned} P_{\text{fuse}} &= p_1 + p_2 + \cdots + p_k \\ &= (p_1^{\text{pre}} + p_2^{\text{pre}} + \cdots + p_k^{\text{pre}}) + body + (p_1^{\text{suf}} + p_2^{\text{suf}} + \cdots + p_k^{\text{suf}}) \\ &= P^{\text{pre}} + body + P^{\text{suf}}. \end{aligned} \tag{2}$$

3.3 Ensemble Learning Based Fusion

Although the syntax based fusion approach is simple and sufficient, it is restricted by the form of text descriptions which may cause ambiguity in the fused prompt during processing. Moreover, the fused prompts are normally too long that the model could lose proper attention to the key information, resulting in extremely unstable performance (results shown in Sect. 4.2).

Therefore, in this subsection, we further explore fusion approaches based on ensemble learning. More specifically, the VLM outputs a set of candidate region proposals C_i for each prompt p_i, and these candidates carry more multi-dimensional information than prompts. We find in our preliminary experiments that direct concatenation of the candidates is not satisfactory and effective, since simply integration hardly screens out the bad predictions. In addition, the candidate, e.g., $c_{ij} \in C_i$, carries richer information that can be further utilized, such as central coordinate x_j and y_j, region size w_j and h_j, category label, and prediction confidence score. Therefore, we consider step-wise clustering mechanisms using the above information to screen out the implausible candidates based on clustering ensemble learning [3].

Another observation in our preliminary experiments is that most of the candidates distribute near the target if the prompt description matches better with the object. Moreover, the candidate regions of inappropriate size containing too much background or only part of the object should be abandoned directly. As such, we consider clustering the center coordinate (x_j, y_j) and region size (w_j, h_j) respectively to filter out those candidates with the wrong location and size.

This step-wise clustering with the aid of different features embodies a typical ensemble learning idea. Therefore, we propose a method called Ensemble Guided Fusion based on semi-clustering ensemble, as detailed in Fig. 1 (b). There are four sub-modules in our approach, where the location cluster f_{loc} and size cluster f_{size} discard the candidates with large deviations and abnormal sizes. Then, in the prediction corrector f_{correct}, we utilize the voting mechanism to select the remaining candidates with appropriate category tags and relatively high prediction confidence. After the first three steps of processing, the remaining candidates $C^{'}$ originated from each prompt can be written as:

$$C^{'} = C \cdot f_{\text{loc}}(C) \cdot f_{\text{size}}(C) \cdot f_{\text{correct}}(C). \tag{3}$$

The remaining candidates are then transferred to the integration module for being integrated into the final fused result C_{fuse} that is mutually independent:

$$C_{\text{fuse}} = \bigcup C^{'}. \tag{4}$$

Besides, we also propose three fusion strategies to recluster candidates in different ways before executing ensemble guided fusion, i.e., fusing the multiple prompts equally, by category, and by attribute. Compared to the first strategy, fusing by category and by attribute both have an additional step of reorgnization. Candidates whose prompts belong to the same category or have identical attributes will share the similar distribution. Accordingly, we rearrange these candidates C_i into a new set C for the subsequent fusion process.

4 Experiments and Results

4.1 Experimental Settings

We collect three public medical image datasets across various modalities including skin lesion detection dataset ISIC 2016 [2], polyp detection dataset CVC-300 [21], and blood cell detection dataset BCCD to validate our proposed approach for zero-shot medical lesion detection. For the experiments, we use the GLIP-T variant [11] as our base pre-trained model and adopt two metrics for the grounding evaluation, including Average Precision (AP) and AP50. More details on the dataset and implementation are described in the appendix.

Table 1. Our approaches v.s. single prompts and other fusion methods.

Prompt	Method	ISIC 2016		CVC-300		BCCD	
		AP	AP50	AP	AP50	AP	AP50
Single	GLIP [11]	10.5	20.0	29.8	37.9	8.9	18.4
		11.3	22.7	16.8	21.7	12.2	23.1
		13.6	25.5	20.2	30.3	9.6	18.2
Multiple	NMS [14]	12.0	20.6	27.9	37.9	11.9	21.4
	Soft-NMS [1]	18.8	30.3	24.1	31.2	11.9	21.4
	WBF [20]	1.16	5.37	3.27	9.40	1.22	4.75
	Concatenation	16.9	27.4	21.5	27.8	11.6	20.6
	Syntax based fusion	13.9	24.1	10.0	16.4	12.8	25.4
	Ours	**19.8**	**30.9**	**36.1**	**47.9**	**15.8**	**32.6**

4.2 Results

This section demonstrates that our proposed ensemble guided fusion approach can effectively benefit the model's performance.

The Proposed Approach Achieves the Best Performance in Zero-Shot Lesion Detection Compared to Baselines. To confirm the validity of our method, we conduct extensive experiments under the zero-shot setting and include a series of fusion baselines: Concatenation, Non-Maximum Suppression (NMS) [14], Soft-NMS [1] and Weighted Boxes Fusion (WBF) [20] for comparisons. As illustrated in Table 1, our ensemble guided fusion rivals the GLIP [11] with single prompt and other fusion baselines across all datasets. The first three rows in Table 1 represent the results of single prompt by only providing shape, color, and location information, respectively. Furthermore, we conduct a comparison between YOLOv5 [17] and our method on CVC-300 under 10-shot settings. Table 2 shows that our method outperforms YOLOv5, which indicates fully-supervised models such as YOLO may not be suitable for medical scenarios where a large labeled dataset is often not available. In addition, we utilize the Automatic Prompt Engineering (APE) [25] method to generate prompts. These prompts give comparable performance to our single prompt and can be still be improved by our fusion method. And the details are described in the appendix.

Table 2. Comparison with YOLOv5.

Method	Training mAP	Test mAP
YOLOv5	16.2	6.4
Ours	**60.8**	**50.2**

Table 3. Zero-shot v.s. 10-shot results.

Dataset	ISIC 2016	CVC-300
Zero-shot	19.8	36.1
10-shot	**38.2**	**50.2**

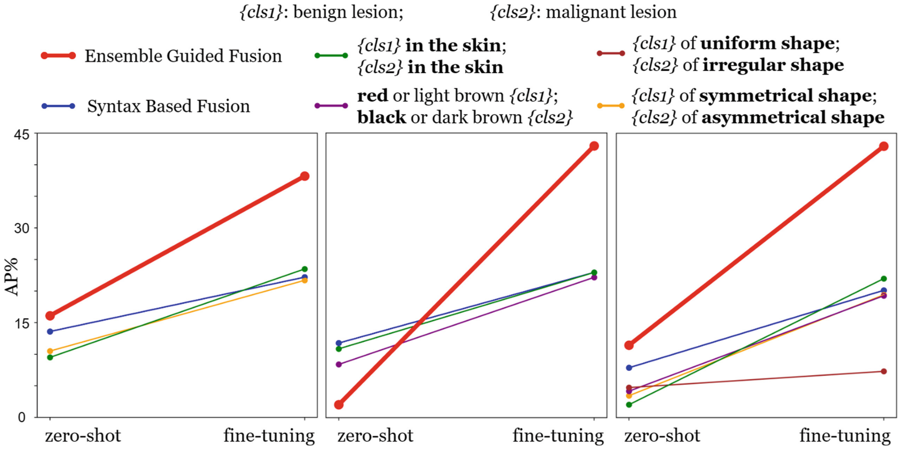

Fig. 2. Fine-tuning v.s. zero-shot results on the ISIC 2016 dataset.

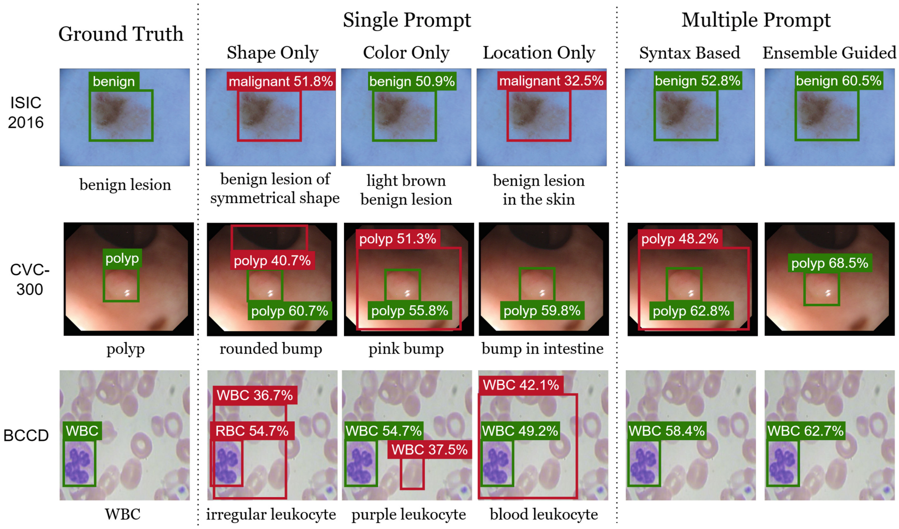

Fig. 3. Comparisons of test results between before and after multi-prompt fusion under zero-shot settings. Here we present part of the single prompts used in the experiments for illustration. The misclassification problem in some of the single prompts is corrected (i.e., malignant to benign) on the first dataset. For all datasets, the candidate boxes are more precise and associated with higher confidence scores.

Fine-Tuned Models Can Further Improve the Detection Performance. We conduct 10-shot fine-tuning experiments as a complement, and find the performance greatly improved. As shown in Table 3 and Fig. 2, with the same group of multiple prompts, the accuracy of fine-tuned model has increased almost twice as much as that of zero-shot, further demonstrating the effectiveness of our

Table 4. Results with different fusion strategies.

Dataset Strategy	ISIC 2016		CVC-300		BCCD	
	AP	AP50	AP	AP50	AP	AP50
Equally	16.8	25.2	30.8	40.4	12.5	21.6
Category	13.2	20.4	30.8	40.4	15.3	24.9
Attribute	19.8	30.9	36.1	47.9	15.8	32.6

method in both settings. Therefore, we can conclude that the pre-trained GLIP model has the ability to learn a reasonable alignment between textual and visual modalities in medical domains.

Visualizations. Figure 3 shows the visualization of the zero-shot results across three datasets. Syntax based fusion sometimes fails to filter out unreasonable predictions because these regions are generated directly by the VLM without further processing and eventually resulting in unstable detection performance. On the contrary, our approach consistently gives a better prediction that defeats all single prompts with a relatively proper description, yet syntax based fusion relies too much on the format and content of inputs, which results in great variance and uninterpretability. The step-wise clustering mechanism based on ensemble learning enables our method to exploit multi-dimensional information besides visual features. In addition, the key components in our proposed approach are unsupervised, which also enhances stability and generalization.

Comparison Among Fusion Strategies. In this work, we not only provide various solutions to fuse multiple prompts but also propose three fusion strategies to validate the generalization of ensemble-guided fusion. As shown in Table 4, we present the results obtained with three different fusion strategies: equally, by category, and by attribute. The first strategy is to process each prompt equally, which is the most convenient and suitable in any situation. Fusing prompts by category is specifically for multi-category datasets to first gather the prompts belonging to the same category and make further fusion. Similarly, Fusing by attribute is to fuse the candidates, whose prompts are describing the same attribute. The strategy of fusing by attribute, outperforms the other ones, due to the fact that candidates with the same attribute share a similar distribution, which is prone to obtain a more reasonable cluster. On the contrary, it is possible to neglect this distribution when fusing each prompt equally.

Ablation Study. As shown in Table 5, we perform ablation studies on three datasets. Our approach has three key components, i.e., location cluster, size cluster and prediction corrector. The location cluster filters out the candidates with severe deviation from the target. The size cluster removes those abnormal ones.

Table 5. Ablation for key components in our proposed approach.

Components			ISIC 2016		CVC-300		BCCD	
Location Cluster	Size Cluster	Prediction Corrector	AP (%)	AP50 (%)	AP (%)	AP50 (%)	AP (%)	AP50 (%)
✓			13.9	26.4	24.2	30.1	10.9	20.2
	✓		10.4	19.5	23.9	29.5	11.9	21.4
		✓	13.8	24.8	29.5	37.3	9.3	17.5
✓	✓		16.9	27.4	30.6	41.7	15.1	31.3
✓	✓	✓	**19.8**	**30.9**	**34.1**	**45.7**	**15.8**	**32.6**

Finally, the prediction corrector further eliminates the candidates that cause low accuracy. The results show that when combining the above three components, the proposed approach gives the best lesion detection performance, suggesting that all components are necessary and effective in the proposed approach.

5 Conclusion

In this paper, we propose an ensemble guided fusion approach to leverage multiple text descriptions when tackling the zero-shot medical lesion detection based on vision-language models and conduct extensive experiments to demonstrate the effectiveness of our approach. Compared to a single prompt that typically requires exhaustive engineering and designation, the multiple medical prompts provide a flexible way of covering all key information that help with lesion detection. We also present several fusion strategies for better exploiting the relationship among multiple prompts. One limitation of our method is that it requires diverse prompts for effective clustering of the candidates. However, with the help of other prompt engineering methods, the limitation can be relatively alleviated.

References

1. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-nms-improving object detection with one line of code. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5561–5569 (2017)
2. Codella, N.C., et al.: Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), pp. 168–172. IEEE (2018)
3. Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q.: A survey on ensemble learning. Front. Comput. Sci. **14**, 241–258 (2020)
4. Gao, P., et al.: Clip-adapter: better vision-language models with feature adapters. arXiv preprint arXiv:2110.04544 (2021)
5. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
6. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)

7. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021)
8. Jensen, J.D., Elewski, B.E.: The ABCDEF rule: combining the "ABCDE Rule" and the "Ugly Duckling Sign" in an effort to improve patient self-screening examinations. J. Clin. Aesthetic Dermatol. **8**(2), 15 (2015)
9. Jiang, C., Wang, S., Liang, X., Xu, H., Xiao, N.: ElixirNet: relation-aware network architecture adaptation for medical lesion detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11093–11100 (2020)
10. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, pp. 12888–12900. PMLR (2022)
11. Li, L.H., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10965–10975 (2022)
12. Li, N., Jiang, Y., Zhou, Z.-H.: Multi-label selective ensemble. In: Schwenker, F., Roli, F., Kittler, J. (eds.) MCS 2015. LNCS, vol. 9132, pp. 76–88. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20248-8_7
13. Li, N., Zhou, Z.-H.: Selective ensemble of classifier chains. In: Zhou, Z.-H., Roli, F., Kittler, J. (eds.) MCS 2013. LNCS, vol. 7872, pp. 146–156. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38067-9_13
14. Neubeck, A., Van Gool, L.: Efficient non-maximum suppression. In: 18th international conference on pattern recognition (ICPR 2006), vol. 3, pp. 850–855. IEEE (2006)
15. Qin, Z., Yi, H., Lao, Q., Li, K.: Medical image understanding with pretrained vision language models: a comprehensive study. arXiv preprint arXiv:2209.15517 (2022)
16. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
17. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
19. Sarcar, M., Rao, K., Narayan, K.: Computer aided design and manufacturing. PHI Learning (2008). https://books.google.co.jp/books?id=zXdivq93WIUC
20. Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: Ensembling boxes from different object detection models. Image Vis. Comput. **107**, 104117 (2021)
21. Vázquez, D., et al.: A benchmark for endoluminal scene segmentation of colonoscopy images. J. Healthcare Eng. **2017** (2017)
22. Wei, S., Li, Z., Zhang, C.: Combined constraint-based with metric-based in semi-supervised clustering ensemble. Int. J. Mach. Learn. Cybern. **9**, 1085–1100 (2018)
23. Zhang, M.L., Zhou, Z.H.: Exploiting unlabeled data to enhance ensemble diversity. Data Min. Knowl. Disc. **26**, 98–129 (2013)
24. Zhang, P., et al.: VinVL: revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5579–5588 (2021)
25. Zhou, Y., et al.: Large language models are human-level prompt engineers (2023)