Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 21, No. 1, 2026

Structure Enhancement and Cross-Modal Alignment for Open-Vocabulary Semantic Segmentation

Jiawei Bai

School of Taiyuan Normal University, Shanxi, 030600, China

Abstract: This paper proposes a structure-enhanced cross-modal alignment method for open-vocabulary semantic segmentation. Existing methods mostly rely on CLIP's image-level vision-language alignment capability, but CLIP visual features remain insufficient for modeling fine-grained spatial information such as boundaries, textures, and region structures. Moreover, relying solely on semantic alignment between images and text categories makes it difficult to model fine-grained correspondences between visual and textual information. To address these issues, we design a DINO Structure Enhancement Module and a Cross-Modal Alignment Module (CCMA). The DINO Structure Enhancement Module introduces a parameter-frozen DINO model to extract structural priors and adaptively enhance CLIP visual features, thereby producing structure-aware visual features. CCMA jointly models global visual semantics, local region features, and text semantic prototypes to mine fine-grained vision-language consistency at the region level, thereby strengthening the correspondence between image regions and textual semantics. Experimental results demonstrate that the proposed method effectively improves open-vocabulary semantic segmentation performance.

Keywords: Open-Vocabulary Semantic Segmentation; Cross-modal Alignment; Structure Enhancement.

1. Introduction

Semantic segmentation aims to assign a semantic label to each pixel in an image and is a fundamental task in computer vision . Conventional semantic segmentation methods are typically trained and evaluated under a closed-set setting, where the categories in the training and test sets are assumed to be identical. Although these methods have achieved strong performance, their generalization ability is heavily constrained by the fixed set of annotated training categories and is limited when novel categories appear in real-world scenarios. Therefore, enabling models to segment unseen categories according to textual category prompts has become an important research direction, leading to the development of open-vocabulary semantic segmentation.

The rapid development of vision-language models has provided an effective foundation for open-vocabulary semantic segmentation. CLIP maps images and texts into a shared semantic embedding space through large-scale image-text contrastive learning, thereby showing strong open-vocabulary generalization ability. However, CLIP is mainly trained for image-level vision-language matching rather than pixel-level dense prediction. As a result, CLIP visual features are limited in modeling fine-grained spatial information such as boundaries, textures, and region structures. Meanwhile, open-vocabulary semantic segmentation requires stable cross-modal correspondences between visual regions and textual categories. Existing methods mostly rely on image-level vision-language alignment, while the fine-grained correspondences between local visual regions and text categories are still insufficiently modeled. This may lead to unstable local region matching, blurred boundaries, and category confusion in complex scenes.

To address these limitations, we propose an open-vocabulary semantic segmentation method that combines DINO-based structure enhancement with cross-modal alignment. The proposed method consists of two key modules: a DINO Structure Enhancement Module and a Cross-Modal Alignment Module (CCMA). The DINO Structure Enhancement Module uses a parameter-frozen DINO model to extract structural priors from the input image and adaptively enhances CLIP visual features through spatial gating, thereby supplementing fine-grained information such as boundaries, textures, and region structures. CCMA jointly models global visual semantics, local region features, and text semantic prototypes. It uses global semantics to guide the reliability modeling of local regions and further establishes fine-grained correspondences between local regions and text semantic prototypes, thereby enhancing cross-modal matching in open-vocabulary scenarios.

The two modules address the limitations of CLIP-based methods from complementary perspectives. The DINO Structure Enhancement Module improves the structural representation of visual features, while CCMA refines image-text alignment into local region–semantic prototype correspondences. Together, they enhance both visual structure modeling and fine-grained cross-modal alignment for open-vocabulary semantic segmentation.

The main contributions of this paper are summarized as follows:

(1) We propose a DINO Structure Enhancement Module to address the insufficient pixel-level structural modeling in open-vocabulary semantic segmentation. By introducing a parameter-frozen DINO model to extract image structural priors, the module adaptively enhances CLIP visual features and improves the representation of fine-grained information such as boundaries, textures, and region structures.

(2) We propose a global-local collaborative fine-grained Cross-Modal Alignment Module (CCMA). By jointly modeling global visual semantics, local region features, and category semantic prototypes, CCMA refines image-text alignment into local region–semantic prototype correspondences, thereby enhancing fine-grained cross-modal alignment in open-vocabulary semantic segmentation.

(3) Experiments on multiple open-vocabulary semantic segmentation benchmarks demonstrate that the proposed

method consistently improves over the baseline, validating the effectiveness and complementarity of the DINO Structure Enhancement Module and CCMA.

2. Related Work

2.1. Semantic Segmentation

Semantic segmentation aims to assign a semantic label to each pixel in an image and is one of the fundamental tasks in computer vision. Existing semantic segmentation methods can be broadly divided into convolutional neural network-based methods and Transformer-based methods. Representative CNN-based methods include FCN, PSPNet, and DeepLab, which improve dense prediction through fully convolutional architectures, pyramid context aggregation, and atrous convolution, respectively. More recently, Transformer-based methods such as SegFormer and Mask2Former have introduced self-attention mechanisms to enhance global context modeling and have achieved strong performance on various segmentation benchmarks.

Despite their success, these methods are generally designed under a closed-set assumption, where the training and test categories are identical. As a result, they struggle to generalize to novel categories that are not observed during training. This limitation has motivated increasing research interest in open-vocabulary semantic segmentation, which aims to segment unseen categories with the help of textual category descriptions.

2.2. Open-Vocabulary Semantic Segmentation

Open-vocabulary semantic segmentation (OVSS) aims to perform pixel-level segmentation of both seen and unseen categories by leveraging textual category prompts. The rapid development of vision-language models, especially CLIP, has provided a strong foundation for this task. Existing OVSS methods can be roughly categorized into two-stage methods and end-to-end methods.

Two-stage methods usually first generate class-agnostic masks or candidate regions and then classify these regions with a vision-language model. Representative methods include OpenSeg and OVSeg. OpenSeg leverages image-level labels to scale open-vocabulary segmentation, while OVSeg adopts mask-adapted CLIP to reduce the gap between image-level vision-language pretraining and mask-level segmentation. These methods effectively exploit the open-vocabulary recognition ability of vision-language models, but they often depend on additional mask or region proposal processes. Their performance is therefore sensitive to the quality of candidate masks and may involve considerable computational cost.

End-to-end methods instead build image-text correspondences directly at the feature level. DenseCLIP extends CLIP from image-text matching to dense pixel-text prediction with context-aware prompting. ZegFormer decouples zero-shot semantic segmentation into class-agnostic mask generation and semantic classification. CAT-Seg further formulates OVSS as a cost aggregation problem between image and text features, improving image-text matching through spatial and class-wise aggregation. Although these methods improve the adaptability of CLIP to dense prediction tasks, they still largely rely on image-level semantic alignment. As a result, CLIP visual features may remain insufficient for modeling fine-grained spatial details such as boundaries, textures, and region structures. Moreover, the correspondences between local visual regions and text categories are still not fully explored.

2.3. Structural Priors and Cross-Modal Alignment

To improve visual representations for pixel-level prediction, recent studies have explored the use of structural priors from self-supervised vision models. DINO and DINOv2 have shown strong capability in capturing boundaries, region structures, and fine-grained visual structures. These properties make self-supervised visual features useful complements to CLIP features, which are mainly learned from image-level vision-language supervision. For example, CLIP-DINOiser incorporates localization priors from self-supervised features into CLIP visual representations to improve spatial localization in open-vocabulary segmentation. Talk2DINO further bridges DINOv2 visual features with language semantics to enhance spatial perception in open-vocabulary segmentation.

Cross-modal alignment is another key issue in open-vocabulary semantic segmentation. Existing works have explored different ways to build correspondences between visual regions and textual semantics. CLIPSeg introduces a prompt-driven segmentation decoder to incorporate text semantics into pixel-level prediction. GroupViT learns visual grouping under text supervision, allowing image regions to emerge with semantic meanings. OVSegmentor further learns open-vocabulary segmentation from natural language supervision by organizing the relationship between pixels and semantic representations. These methods demonstrate the importance of cross-modal alignment for OVSS.

However, existing methods still have limitations in jointly exploiting global semantic context and local region-level discriminative information. Global semantics provide category-level contextual cues, while local regions contain fine-grained visual evidence for segmentation. Insufficient interaction between these two levels limits the ability of existing models to capture precise region-text correspondences in complex scenes. Motivated by this observation, our method improves OVSS from two complementary perspectives. First, we introduce a frozen DINO model to provide structural priors and enhance the structure-aware representation of CLIP visual features through spatial gating. Second, we propose a global-local collaborative cross-modal alignment module that jointly models global visual semantics, local region features, and text semantic prototypes to establish finer-grained vision-language correspondences at the region level.

3. Methods

3.1. Overall Framework

Fig. 1 illustrates the overall framework proposed in this paper. The proposed framework is built upon CAT-Seg and consists of two core components: a DINO structure enhancement module and a cross-modal alignment module (CCMA). First, the input image is processed through the CLIP image encoder to obtain visual features, while simultaneously being fed into a frozen DINO model to extract structural priors. The DINO structural enhancement module generates a spatial gating map based on the structural prior and performs adaptive enhancement on the CLIP visual features to supplement boundary and regional structural information. Concurrently, the CCMA module jointly models the

relationships between global visual semantics, local regional features, and textual semantic prototypes based on CLIP visual and textual features, generating region-consistency-guided cross-modal alignment scores. Subsequently, the model utilises the DINO-enhanced visual features and textual features to model image-text correlations, and integrates the alignment information output by the CCMA into the matching process to produce the final open-vocabulary segmentation prediction.

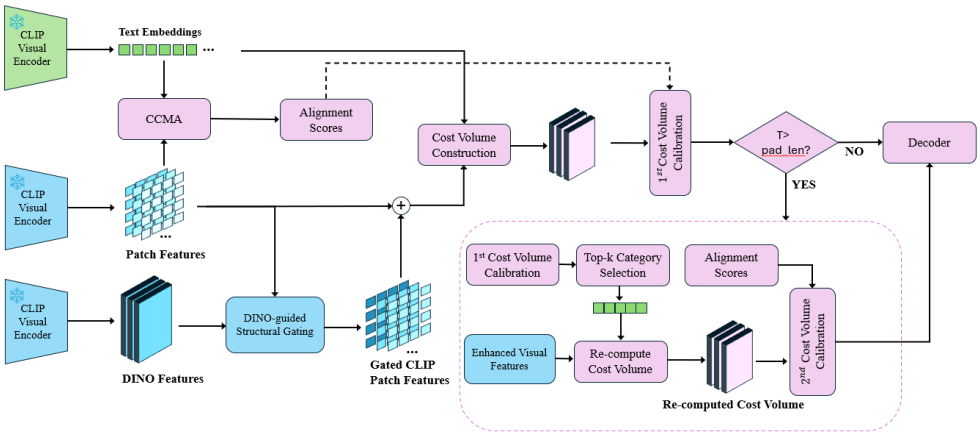

Fig. 1 Overall architecture of the model

3.2. DINO Structural Enhancement Module

To address the shortcomings of CLIP visual features in representing local structures, this paper introduces a frozen DINO model as a structural prior extractor. As shown in Fig. 2, the input image first undergoes pre-processing consistent with CLIP, including normalisation and scaling to a fixed resolution; subsequently, the image is fed into both the CLIP image encoder and the frozen DINO model, yielding CLIP visual features and DINO structural features respectively. The DINO branch maintains parameter freezing during both training and inference, serving solely as an external structural prior for feature enhancement. Let the visual features extracted by the CLIP image encoder be $F_v \in \mathbb{R}^{C\times H\times W}$, where C, H and W represent channel, height and width respectively. The structural features extracted by the DINO model are $F_d \in \mathbb{R}^{C_d\times H_d\times W_d}$

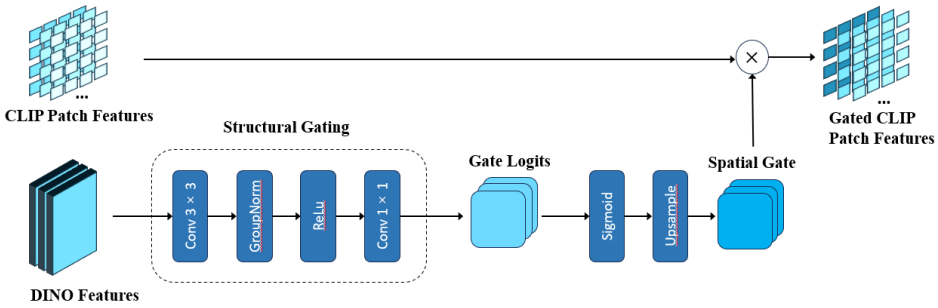

Fig. 2 DINO Structure Enhancement Module

During the structural enhancement process, the DINO structural feature F_d is first mapped using a gating unit. This gating unit transforms the DINO feature through a sequence of operations, including convolution, normalisation and non-linear activation, to generate a single-channel spatial gating response:

$$M = \varphi(F_d),$$

where, φ $(\cdot)$denotes the gate-mapping function, Subsequently, M is scaled and passed through a sigmoid activation function, then upsampled to the same spatial resolution as the CLIP visual features, yielding the final spatial gating map, denoted as $G \in \mathbb{R}^{1\times H\times W}$ By applying multiplicative modulation to visual features using a spatial gating map, we obtain enhanced features:

$$\widehat{F_v} = F_v \odot (1 + \alpha G),$$

where α is the learnable modulation coefficient, $\odot$ denotes element-wise multiplication. To maintain the stability of the original semantic representation, this paper further employs a weighted residual fusion strategy to combine the original visual features with the enhanced features:

$$F_v' = (1 - \gamma)F_v + \gamma\widehat{F_v},$$

where γ is the learnable fusion coefficient. The resulting enhanced visual features$F_v^{'} \in \mathbb{R}^{C\times H\times W}$ It retains the same dimensions as the original visual features and is used for subsequent segmentation.

3.3. Cross-Modal Alignment Module (CCMA)

As shown in Figure 3, CCMA aims to enhance the fine-grained correspondence between visual features and textual semantics at the region level. Unlike direct global image-text matching, CCMA jointly uses global visual semantics and local region features. The global visual representation provides scene-level semantic context, while local region features provide fine-grained visual evidence for matching with text semantic prototypes. Through this global-local collaborative modeling strategy, CCMA refines coarse image-text alignment into local region–semantic prototype correspondences. Let the visual features be $F_v \in \mathbb{R}^{C\times H\times W}$, and let the text features are $F_t \in \mathbb{R}^{T\times P\times C}$, where T denotes the number of classes, P denotes the number of prompt templates, and C denotes the feature dimension.

In the visual branch, the visual features F_v are first

normalized. A global visual feature are obtained via global average pooling. f_g. Meanwhile, adaptive average pooling is used to divide the visual features into r × r local region features, yielding a set of local features $\{f_i\}_{i=1}^{N}$, $N = r^2$, $f_i \in \mathbb{R}^C$ Subsequently, by calculating the consistency between global visual features and the features of each local region, the regional reliability weights are obtained w_i This weight captures the consistency between the local region and the semantic context of the entire image. Local regions that are more consistent with the global semantic context typically contain clearer cues for class identification and should therefore be assigned a higher contribution in subsequent region-text matching.

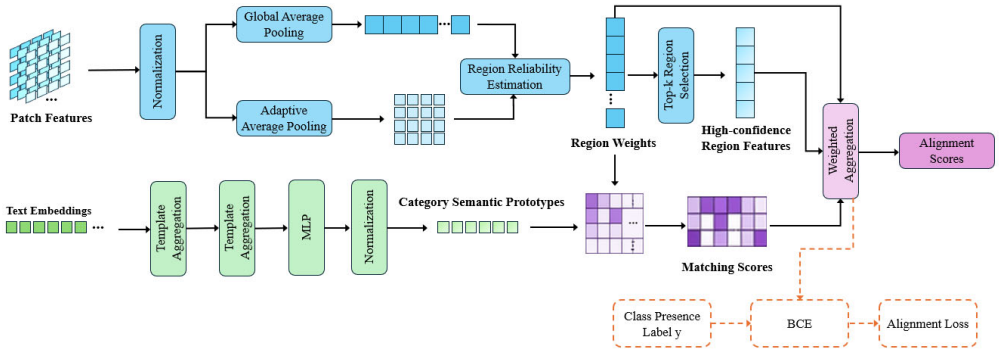

Fig. 3 Cross-modal alignment module

$$w_i = \frac{\psi(f_g, f_i)}{\sum_{j=1}^{N} \psi(f_g, f_j)}$$

where, $\psi(\,,\,)$ denotes the similarity function, and w_i reflects the contribution of the i-th local region to the subsequent alignment. Based on this weighting, the top-k high-confidence regions are further selected for subsequent semantic matching. In the text branch, the text features of multiple templates within the same category are first aggregated along the template dimension to obtain the category text representation:

$$\bar{t}_t = \frac{1}{P}\sum_{p=1}^{P} F_t(t,p), \qquad \bar{t}_t \in \mathbb{R}^C$$

Then, a lightweight mapping function is applied to construct text semantic prototypes:

$$c_t = Norm\left(\bar{t}_t + \Delta(\bar{t}_t)\right), \qquad c_t \in \mathbb{R}^{\mathbb{C}}$$

where $\Delta(\cdot)$ denotes a lightweight mapping function, and Norm$(\cdot)$ denotes a normalization operation. In this way, category representations can be enhanced while preserving the original semantic information. After obtaining high-confidence region features and class semantic prototypes, this paper further calculates the matching score between the two. Let the i-th high-confidence region feature be f_i and the t-th class semantic prototype be c_t; then the matching score between the region and the class is expressed as:

$$s_{i,t} = \tau\langle f_i, c_t\rangle, \qquad i = 1, \ldots, k$$

where, $\langle\cdot,\cdot\rangle$ denotes the inner product and τ is a learnable temperature coefficient. The final alignment score for class t is obtained by weighted aggregation over the selected local regions:

$$s_t = \sum_{i=1}^{k} w_i \; s_{i,t}$$

The final alignment score is obtained by weighted aggregation of the matching relationships between multiple high-confidence local regions and text semantic prototypes. Compared with image-text matching that relies only on global visual features, this score can better reflect the fine-grained consistency between local visual regions and open-vocabulary textual semantics. During training, pixel-level annotations are used to derive image-level semantic presence labels, which provide auxiliary supervision for the alignment scores produced by CCMA and encourage the model to learn more stable region-text correspondences. The auxiliary alignment loss is defined as:

$$\mathcal{L}_{align} = \mathrm{BCE}(s, y)$$

where y denotes the image-level class presence label derived from the segmentation annotation, BCE$(\cdot)$ denotes the binary cross-entropy loss. During the forward matching stage, alignment scores are used to enhance the image-text relevance. Let the image-text relevance be $Corr \in \mathbb{R}^{P\times T\times H\times W}$ It is calculated based on the correlation between visual features and textual features. The alignment scores are first normalized and then transformed into an alignment guidance factor:

$$\widetilde{s_t} = \frac{s_t - \mu_s}{\sigma_s}$$

$$g_t = 1 + \beta \tanh(\widetilde{s_t})$$

where $\widetilde{s_t}$ denotes the normalized alignment score, and β is a fixed modulation coefficient. The alignment guidance factor is incorporated into the image-text correlation modeling process:

$$Corr'_{p,t,h,w} = g_t \cdot Corr_{p,t,h,w}$$

In this way, text categories that are more consistent with local region semantics obtain more stable responses in subsequent matching, thereby enhancing the correspondence between image regions and textual semantics in open-vocabulary scenarios.

Furthermore, in large-vocabulary scenarios, the number of candidate categories can be very large. To control computational cost and reduce the influence of numerous low-response categories, the model selects candidate categories according to image-text correlation responses and reconstructs the image-text correlation over the selected category set. The alignment information generated by CCMA is also applied to the filtered category set, ensuring consistency between the original and filtered matching processes.

4. Experiments

4.1. Datasets and Evaluation Metrics

We train the proposed model on COCO-Stuff and evaluate

it directly on several downstream open-vocabulary semantic segmentation benchmarks, including ADE20K, PASCAL VOC, and PASCAL-Context. COCO-Stuff contains approximately 118K training images with pixel-level annotations covering 171 semantic categories. For evaluation, ADE20K provides two commonly used settings: A-150 with 150 categories and A-847 with 847 categories. PASCAL VOC is evaluated under the PAS-20 and PAS-20b settings. PASCAL-Context extends the original PASCAL VOC dataset and provides two evaluation settings: PC-59 and PC-459, containing 59 and 459 categories, respectively.

We adopt mean Intersection over Union (mIoU) as the evaluation metric. For class c, the IoU is defined as:

$$mIoU = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_c}{TP_c + FP_c + FN_c}$$

where TP_c denotes the number of pixels correctly predicted to belong to class c, FP_c denotes the number of pixels incorrectly predicted to belong to class c, FN_c denotes the number of pixels that actually belong to class c but were not correctly predicted as such, and K denotes the total number of classes.

4.2. Experimental Details

This paper follows the standard experimental setup for open-vocabulary semantic segmentation, training models on the COCO-Stuff dataset and testing them directly on downstream benchmark datasets such as ADE20K, PASCAL VOC and PASCAL-Context. CAT-Seg is used as the baseline model; the implementation is based on PyTorch, and training was performed on two NVIDIA RTX 5090 GPUs. During training, the CLIP and DINO models were fully frozen. AdamW was used as the optimiser. The total loss during the training phase consisted of the semantic segmentation loss and an auxiliary alignment loss, with the weight of the auxiliary alignment loss set to 0.1. The training batch size was set to 4, the learning rate to 0.0002, and the model underwent a total of 110,000 training iterations.

4.3. Main Results

Table 1 compares the proposed method with existing open-vocabulary semantic segmentation methods on multiple benchmark datasets. Overall, our method consistently improves over the CAT-Seg baseline, demonstrating the effectiveness of the proposed structure enhancement and cross-modal alignment design.

Specifically, on A-150, PAS-20, and PAS-20b, the proposed method achieves higher mIoU than the baseline, indicating that the introduced modules can provide stable improvements under relatively common open-vocabulary evaluation settings. In addition, our method also maintains consistent gains on more challenging large-vocabulary benchmarks such as A-847 and PC-459. These datasets contain a larger number of fine-grained categories, making the matching between image regions and textual categories more difficult. The results suggest that the global-local cross-modal alignment strategy is beneficial for modeling semantic correspondences in complex category spaces.

Combined with the ablation studies and qualitative results, the improvements can be attributed to two complementary factors: the DINO Structure Enhancement Module improves the structural representation of visual features, while CCMA enhances fine-grained region-text correspondence. Overall, the results validate the effectiveness of the proposed method in both structure-aware visual representation and fine-grained cross-modal alignment.

Table 1. Quantitative evaluation on standard benchmarks.

Model	VLM	Backbone	Training Dataset	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20b
SPNet [23]	–	ResNet-101	PASCAL VOC	–	–	–	24.3	18.3	–
ZS3Net [24]	–	ResNet-101	PASCAL VOC	–	–	–	19.4	38.3	–
LSeg [25]	CLIP ViT-B/32	ResNet-101	PASCAL VOC-15	–	–	–	–	47.4	–
ZegFormer [26]	CLIP ViT-B/16	ResNet-101	COCO-Stuff-156	4.9	9.1	16.9	42.8	86.2	62.7
ZSseg [27]	CLIP ViT-B/16	ResNet-101	COCO-Stuff	7.0	–	20.5	47.7	88.4	–
OpenSeg [7]	ALIGN	ResNet-101	COCO Panoptic	4.4	7.9	17.5	40.1	–	63.8
OVSeg [8]	CLIP ViT-B/16	ResNet-101c	COCO-Stuff	7.1	11.0	24.8	53.3	92.6	–
ZegCLIP [28]	CLIP ViT-B/16	–	COCO-Stuff-156	–	–	–	41.2	93.6	–
SAN [29]	CLIP ViT-B/16	–	COCO-Stuff	10.1	12.6	27.5	53.8	94.0	–
SED [30]	ConvNeXt-B	–	COCO-Stuff	11.4	18.6	31.6	57.3	94.4	–
CAT-Seg [11]	CLIP ViT-B/16	–	COCO-Stuff	12.0	19.0	31.8	57.5	94.6	77.3
Ours	CLIP ViT-B/16	–	COCO-Stuff	12.5	19.3	31.9	57.7	95.2	78.2

4.4. Qualitative Results

Fig. 4 presents qualitative comparisons between the proposed method and the CAT-Seg baseline. CAT-Seg tends to produce incomplete object regions, background misclassification, and blurred class boundaries in complex scenes. For example, in animal scenes, CAT-Seg fails to separate foreground objects from background regions consistently, resulting in fragmented object masks. In food scenes, the boundaries between plates, rice, and vegetables are often mixed. In indoor scenes, the segmentation boundaries of objects such as cabinets, refrigerators, and cats are not sufficiently clear.

In contrast, the proposed method produces more complete object regions and clearer boundary responses. This improvement can be attributed to the DINO Structure Enhancement Module, which provides additional structural priors for boundaries, textures, and region consistency. Meanwhile, CCMA enhances the correspondence between local visual regions and text categories by jointly modeling global visual semantics, local region features, and text semantic prototypes. As a result, the proposed method shows better performance in foreground localization, region completeness, and semantic matching consistency.

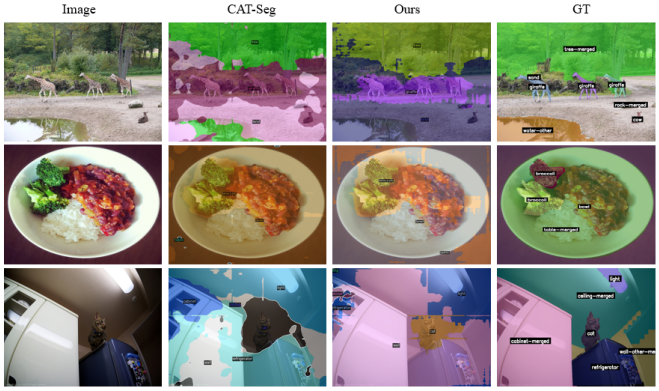

Fig. 4 Qualitative comparison with the CAT-Seg baseline.

4.5. Ablation Studies

To verify the effectiveness of each component, we conduct ablation studies by progressively introducing the DINO Structure Enhancement Module and CCMA into the CAT-Seg baseline.

Table 2. Ablation study of the proposed modules.

DINO	CCMA	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20b
Baseline		12.0	19.0	31.8	57.5	94.6	77.3
√		12.5	18.7	31.9	57.2	94.9	77.7
	√	12.4	19.2	31.6	58.1	95.0	78.2
√	√	12.5	19.3	31.9	57.7	95.2	78.2

As shown in Table 2, introducing the DINO Structure Enhancement Module alone improves performance on multiple datasets. This indicates that the structural priors provided by DINO can compensate for the limitations of CLIP visual features in modeling boundaries, textures, and region structures. When CCMA is introduced alone, the model also achieves clear improvements on datasets such as PASCAL-Context and PASCAL VOC, suggesting that global-local cross-modal alignment helps strengthen the correspondence between local visual regions and text semantic prototypes.

When both modules are introduced simultaneously, the model achieves the best or near-best overall performance. This demonstrates the complementarity of the two modules: the DINO Structure Enhancement Module mainly improves the structural representation of visual features, while CCMA enhances the fine-grained matching between visual regions and textual semantics. Together, they improve open-vocabulary semantic segmentation from both visual representation and cross-modal alignment perspectives.

5. Conclusion

This paper proposes a structure-enhanced cross-modal alignment method for open-vocabulary semantic segmentation. To address the limited structural representation of CLIP visual features, we introduce a parameter-frozen DINO model as an external structural prior and use spatial gating to adaptively enhance CLIP visual features. This design improves the representation of fine-grained spatial information, such as boundaries, textures, and region structures. To further improve cross-modal matching, we propose a global-local collaborative Cross-Modal Alignment Module (CCMA), which jointly models global visual semantics, local region features, and text semantic prototypes. By refining image-text alignment into local region–semantic prototype correspondences, CCMA strengthens fine-grained vision-language consistency in open-vocabulary scenarios.

Experimental results on multiple open-vocabulary semantic segmentation benchmarks show that the proposed method consistently improves over the CAT-Seg baseline. Ablation studies further demonstrate that the DINO Structure Enhancement Module and CCMA are complementary: the former enhances structure-aware visual representation, while the latter improves fine-grained region-text alignment. Together, they improve segmentation performance and generalization ability in open-vocabulary settings.

Nevertheless, the proposed method introduces an additional frozen DINO branch, which brings extra computational overhead during training and inference. In future work, we will explore more lightweight ways to extract structural priors, or distill structural information into compact auxiliary modules to improve efficiency.

Acknowledgements

This work was supported by the Shanxi Provincial Postgraduate Education Innovation Programme (2025SJ411).

References

[1] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, p. 3431-3440.

[2] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 2881-2890.

[3] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence. 2017, 40(4), p. 834-848.

[4] Xie E, Wang W, Yu Z, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers[J]. Advances in neural information processing systems. 2021, 34, p. 12077-12090.

[5] Cheng B, Misra I, Schwing A G, et al. Masked-attention mask transformer for universal image segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 1290-1299.

[6] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PmLR, 2021, p. 8748-8763.

[7] Ghiasi G, Gu X, Cui Y, et al. Scaling open-vocabulary image segmentation with image-level labels[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022, p. 540-557.

[8] Liang F, Wu B, Dai X, et al. Open-vocabulary semantic segmentation with mask-adapted clip[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 7061-7070.

[9] Rao Y, Zhao W, Chen G, et al. Denseclip: Language-guided dense prediction with context-aware prompting[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 18082-18091.

[10] Ding J, Xue N, Xia G S, et al. Decoupling zero-shot semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 11583-11592.

[11] Cho S, Shin H, Hong S, et al. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024, p, 4113-4123.

[12] Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 9650-9660.

[13] Oquab M, Darcet T, Moutakanni T, et al. Dinov2: Learning robust visual features without supervision[J]. arXiv preprint arXiv:2304.07193, 2023.

[14] Wysoczańska M, Siméoni O, Ramamonjisoa M, et al. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024, p. 320-337.

[15] Barsellotti L, Bianchi L, Messina N, et al. Talking to dino: Bridging self-supervised vision backbones with language for

open-vocabulary segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025, p. 22025-22035.

[16] Lüddecke T, Ecker A. Image segmentation using text and image prompts[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 7086-7096.

[17] Xu J, De Mello S, Liu S, et al. Groupvit: Semantic segmentation emerges from text supervision[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 18134-18144.

[18] Xu J, Hou J, Zhang Y, et al. Learning open-vocabulary semantic segmentation models from natural language supervision[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 2935-2944.

[19] Caesar H, Uijlings J, Ferrari V. Coco-stuff: Thing and stuff classes in context[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 1209-1218.

[20] Zhou B, Zhao H, Puig X, et al. Scene parsing through ade20k dataset[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 633-641.

[21] Everingham M, Eslami S M A, Van Gool L, et al. The pascal visual object classes challenge: A retrospective[J]. International journal of computer vision, 2015, 111(1), p. 98-136.

[22] Mottaghi R, Chen X, Liu X, et al. The role of context for object detection and semantic segmentation in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014, p. 891-898.

[23] Xian Y, Choudhury S, He Y, et al. Semantic projection network for zero-and few-label semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019, p. 8256-8265.

[24] Bucher M, Vu T H, Cord M, et al. Zero-shot semantic segmentation[J]. Advances in Neural Information Processing Systems, 2019, 32.

[25] Li B, Weinberger K Q, Belongie S, et al. Language-driven semantic segmentation[J]. arXiv preprint arXiv:2201.03546, 2022.

[26] Ding J, Xue N, Xia G S, et al. Decoupling zero-shot semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 11583-11592.

[27] Xu M, Zhang Z, Wei F, et al. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022, p. 736-753.

[28] Zhou Z, Lei Y, Zhang B, et al. Zegclip: Towards adapting clip for zero-shot semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 11175-11185.

[29] Xu M, Zhang Z, Wei F, et al. Side adapter network for open-vocabulary semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 2945-2954.

[30] Xie B, Cao J, Xie J, et al. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, p. 3426-3436.