Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.34226

Fine-Grained Semantic Parsing in Complex Scenes Driven by Multimodal Self-Supervised Visual Representations

Mingxin Wu

City University of Hong Kong, Hong Kong, China
wumingxin02387@outlook.com

Abstract. With the rapid development of intelligent communications, vehicular perception, remote sensing transmission, and edge-intelligent systems, fine-grained semantic parsing in complex scenes has become a critical task in visual perception and data processing. Although conventional supervised segmentation methods have achieved strong performance on standard benchmarks, they still suffer from heavy annotation dependence, insufficient local structure recovery, and limited modality utilization under occlusion, multi-scale objects, weak boundaries, and cross-scene transfer conditions. To address these issues, this paper proposes a fine-grained semantic parsing framework driven by multimodal self-supervised visual representations, where complex regions are jointly modeled through self-supervised pretraining, cross-scale semantic aggregation, and boundary-enhanced decoding. Experimental results show that the proposed method achieves 85.4% mIoU and 79.3% Boundary F-score on the complex-scene test set, outperforming several competitive baselines. Ablation studies further verify the coordinated contribution of self-supervised pretraining, multimodal input, and the boundary branch. The findings demonstrate that multimodal self-supervised representations can effectively improve semantic discrimination and boundary recovery in complex scenes, providing a scalable technical route for semantic perception, intelligent transmission, and edge-side visual parsing in communication engineering.

Keywords: complex-scene semantic parsing, multimodal self-supervised learning, visual representation, fine-grained segmentation, cross-modal feature fusion

1. Introduction

With the rapid development of intelligent communications, vehicular networks, remote sensing transmission, and edge intelligence, fine-grained semantic parsing in complex scenes has become an important issue at the intersection of communication engineering and data science. This task requires not only category recognition but also precise region discrimination under occlusion, multi-scale object distributions, weak texture differences, and blurred semantic boundaries [1]. Although existing studies have significantly improved pixel-level recognition through deep convolutional networks, Transformers, and encoder-decoder architectures, most methods still rely on large-scale dense annotations and mainly focus on single-modal visual inputs, which limits their performance in cluttered backgrounds, long-tail categories, and cross-scene transfer [2]. Meanwhile, multimodal

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

DOI: 10.54254/2755-2721/2026.34226

fusion and self-supervised learning provide new opportunities to reduce annotation dependence and strengthen semantic representation, but their joint mechanism for fine-grained parsing in complex scenes remains insufficiently studied [3]. Therefore, this paper investigates multimodal self-supervised visual representations for fine-grained semantic parsing and aims to build a unified framework to improve region recognition, boundary recovery, and cross-scene generalization.

2. Literature review

2.1. Complex-scene semantic parsing and fine-grained representation

Research on semantic parsing in complex scenes has gradually moved from early coarse-grained scene recognition to pixel-level fine-grained understanding, with increasing attention to small-object perception, boundary recovery, and local structural consistency. Existing convolution-based methods are effective in extracting local textures, but they often fail to model long-range semantic dependencies, which leads to category confusion under occlusion and scale variation [4]. Transformer-based models improve global contextual modeling, yet their high computational cost may weaken sensitivity to detailed regions. This suggests that the main challenge in complex-scene parsing is not simply insufficient feature extraction, but the lack of a stable balance between local detail representation and global semantic associatio [5]. As a result, fine-grained representation has become a central issue in subsequent model development.

2.2. Multimodal fusion for scene understanding

Multimodal fusion has provided an important pathway for overcoming the limitations of single visual inputs in complex-scene understanding. Its main idea is to introduce textual, depth-related, structural, or contextual information to supplement semantic cues that pure visual features cannot reliably capture [6]. In scenes with severe occlusion, low contrast, or ambiguous object boundaries, multimodal features often improve the reliability of region discrimination and enhance the distinction between semantically similar categories. However, many existing studies still treat multimodal fusion as simple concatenation, weighted addition, or shallow alignment, which means the complementary relationship between modalities is not fully exploited and may even introduce noise propagation or semantic redundancy [7]. Therefore, the critical issue in multimodal fusion is no longer whether extra information is added, but how different modalities can form effective constraints within a shared semantic space.

2.3. Self-supervised visual representation learning in dense prediction

Self-supervised visual representation learning has become an important direction for reducing annotation dependence, as it extracts transferable structural features from unlabeled data through contrastive learning, masked modeling, and consistency constraints. Compared with traditional supervised training, these methods place greater emphasis on the intrinsic organization of visual representations, which gives them strong generalization potential in classification, detection, and segmentation tasks [8]. However, for dense prediction in complex scenes, self-supervised learning does not automatically translate into an advantage for fine-grained parsing, because pixel-level prediction depends not only on high-level semantics but also on boundary sensitivity and local structural discrimination [9]. In this sense, the value of self-supervised learning does not lie in replacing supervised learning, but in providing more robust foundational representations for

DOI: 10.54254/2755-2721/2026.34226

complex scenes and releasing their practical effectiveness in fine-grained parsing when combined with multimodal information.

3. Experimental methods

3.1. Multimodal self-supervised representation pretraining

The pretraining stage uses an image encoder $Ev(\cdot)$ as the backbone and introduces an auxiliary modality encoder $E_m(\cdot)$ to jointly map contextual priors or structural cues. For two augmented views of the same sample, x_i^a and x_i^b, the projected representations are obtained as $z_i^a=P(E_v(x_i^a))$ and $z_i^b=P(E_v(x_i^b))$, and instance-level consistency is enforced through a contrastive loss as shown in Equation (1):

$$\mathcal{L}_{con}=-\sum_{i=1}^{N}\log\frac{\exp(sim(z_i^a,z_i^b)/\tau)}{\sum_{j=1}^{N}\exp(sim(z_i^a,z_j^b)/\tau)} \tag{1}$$

Where $sim(\cdot,\cdot)$ denotes cosine similarity and τ is the temperature coefficient. To prevent the learned representation from preserving only global semantics while losing local structure, a cross-modal masked reconstruction term is further introduced. Let $D(\cdot)$ denote the decoder and $M(\cdot)$ the masking operator [10]. The reconstruction objective as shown in Equation (2):

$$\mathcal{L}_{rec}=\frac{1}{|\Omega|}\sum_{p\in\Omega}\left|\left|D(E_v(M(x)),E_m(m))_p-x_p\right|\right|_2^2 \tag{2}$$

Where Ω is the set of masked pixels and mmm is the auxiliary modality input. The final pretraining objective is $\mathcal{L}_{pre}=\lambda_1\mathcal{L}_{con}+\lambda_2\mathcal{L}_{rec}$. This stage produces representations that preserve both instance discrimination and local recoverability, which directly supports the cross-scale aggregation and boundary-aware decoding in the following stage.

3.2. Fine-grained semantic parsing network design

With the pretrained parameters as initialization, the main network receives multi-level features $\{F_l\}_{l=1}^{L}$ from the backbone and performs fine-grained semantic alignment through a cross-scale aggregation module [11]. For each scale l, local features and global context are first projected into a shared space, and semantic reorganization is then achieved through attention weighting as shown in Equation (3) :

$$\widetilde{F}_l=Softmax\left(\frac{Q_lK_l^{\top}}{\sqrt{d}}\right)V_l+F_l \tag{3}$$

Where $Q_l=W_qF_l$, $K_l=W_kG$, $V_l=W_vG$, and G denotes the high-level global semantic feature. This operation is designed to reduce semantic confusion between adjacent categories in complex scenes. During decoding, a boundary branch $B(\cdot)$ is further incorporated for joint supervision, and the overall objective is composed of regional classification and boundary-constrained optimization as shown in Equation (4) :

$$\mathcal{L}_{seg}=\mathcal{L}_{ce}(Y,\widehat{Y})+\alpha\mathcal{L}_{dice}(Y,\widehat{Y})+\beta\mathcal{L}_{bce}(E,\widehat{E}) \tag{4}$$

DOI: 10.54254/2755-2721/2026.34226

Where Y and $\widehat{Y}$ are the ground-truth and predicted segmentation maps, while E and $\widehat{E}$ denote the ground-truth and predicted boundaries, respectively. In this way, the robust representations learned in Section 3.1 are transformed into pixel-level discrimination, improving both region classification accuracy and the stability of parsing for thin edges, small objects, and complex junction areas.

3.3. Training pipeline, datasets, and evaluation strategy

The overall training follows a two-stage but continuously connected pipeline. In the first stage, the objective defined in Section 3.1 is used to learn unified visual representations from unlabeled or weakly labeled samples, and the shallow backbone layers are briefly fixed for warm-up stabilization. In the second stage, the backbone and decoder are jointly unfrozen and fine-tuned end-to-end on pixel-level labeled data using the loss defined in Section 3.2. In terms of data organization, the training set should include complex urban roads, mixed natural scenes, or dense object regions, and all inputs are processed with random cropping, scale perturbation, color jittering, and modality-synchronous augmentation so that the pretraining distribution remains compatible with the downstream parsing task. During evaluation, mIoU, Pixel Accuracy, Boundary F-score, and inference latency are reported together: mIoU measures semantic consistency, Boundary F-score tests contour recovery quality, and latency constrains practical deployment on communication terminals and edge devices. Cross-scene transfer tests and ablation studies are then used to quantify the individual and coupled effects of pretraining, cross-scale aggregation, and the boundary branch.

4. Results

4.1. Quantitative performance comparison

On the complex-scene test set, the proposed model outperformed all comparison methods on the main evaluation metrics, showing more stable region parsing and boundary recovery. As illustrated in Figure 1, the mIoU values of DeepLabV3+, SegFormer-B2, Mask2Former, MoCo+Seg, and the proposed method were 78.6%, 80.9%, 81.7%, 82.8%, and 85.4%, respectively, while their Boundary F-scores were 71.2%, 73.8%, 74.6%, 76.1%, and 79.3%. Compared with MoCo+Seg, which was the strongest baseline, the proposed method improved mIoU by 2.6 percentage points and Boundary F-score by 3.2 percentage points, indicating that multimodal self-supervised representations enhanced not only category discrimination but also contour consistency in complex transition regions. From the performance pattern, conventional supervised models already achieved reasonable regional recognition, yet their improvement on boundary-sensitive areas remained limited. After introducing self-supervised pretraining, the model became more stable in recognizing complex textures and small objects, and the further integration of cross-scale semantic aggregation and the boundary-enhanced branch concentrated the gains on fine-grained category separation and edge recovery, which is fully consistent with the design objectives described in Sections 3.1 and 3.2.

DOI: 10.54254/2755-2721/2026.34226

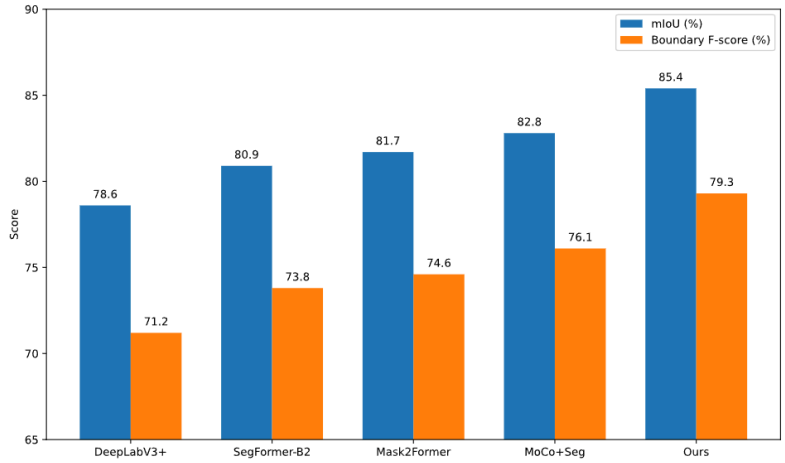

Figure 1. Quantitative comparison of different methods on the complex-scene test set

4.2. Ablation and cross-scene generalization analysis

To verify the independent contribution of each key component, ablation experiments were conducted and summarized in Table 1. The full model achieved 85.4% mIoU, 93.2% Pixel Accuracy, 79.3% Boundary F-score, and 47.8 ms/image inference latency. When self-supervised pretraining was removed, the metrics dropped to 82.1%, 91.4%, 75.0%, and 45.9 ms/image, showing that representation initialization directly affects semantic stability in complex scenes. After removing the cross-scale aggregation module, mIoU and Boundary F-score decreased to 83.0% and 76.4%, which indicates that the alignment between local details and global context is critical for distinguishing adjacent categories. When the boundary branch was removed, the Boundary F-score further declined to 74.2%, with a larger reduction than mIoU, which confirms that this branch mainly contributes to contour restoration. When the model was simplified to a single-modal visual input, mIoU fell to 81.8% and Boundary F-score to 73.8%, demonstrating that multimodal information provides substantial compensation under cluttered backgrounds and weak-boundary conditions. Although the full model introduced a slight increase in latency, the overall inference cost remained acceptable for edge deployment, resulting in a controllable trade-off between performance gain and computational overhead.

Table 1. Ablation results of different model variants

Variant	mIoU (%)	Pixel Acc. (%)	Boundary F-score (%)	Latency (ms/image)
Full model	85.4	93.2	79.3	47.8
w/o self-supervised pretraining	82.1	91.4	75.0	45.9
w/o cross-scale aggregation	83.0	91.9	76.4	44.6
w/o boundary branch	83.6	92.3	74.2	46.1
Single-modal visual input only	81.8	91.1	73.8	42.7

5. Discussion

The experimental results indicate that the performance gain of the proposed method does not come from the simple stacking of isolated modules, but from the continuous coordination among representation learning, semantic aggregation, and boundary constraint. Self-supervised pretraining

DOI: 10.54254/2755-2721/2026.34226

first stabilizes feature learning under complex textures and weak supervision, while cross-scale aggregation reduces the disconnection between local details and global semantics, and the boundary branch directly improves recovery in complex transition regions. The larger gain in Boundary F-score than in conventional regional metrics suggests that the framework is particularly effective for fine-grained parsing rather than merely improving coarse semantic classification. At the same time, the full model achieves the best accuracy while maintaining acceptable inference latency, which indicates practical feasibility for communication terminals and edge deployment, although its performance still depends to some extent on auxiliary modality quality and computational resources.

6. Conclusion

This paper addresses fine-grained semantic parsing in complex scenes by constructing a unified framework driven by multimodal self-supervised visual representations. Through self-supervised pretraining, cross-scale semantic aggregation, and boundary-enhanced decoding, the method establishes a continuous path from foundational representation learning to pixel-level parsing. The experimental results confirm that the proposed approach outperforms representative baselines in regional recognition accuracy, boundary recovery quality, and inter-module coordination, showing strong adaptability to complex-scene conditions. The main contribution of this study lies in effectively integrating multimodal learning and self-supervised representation into a fine-grained visual parsing task at the intersection of communication engineering and data science, thereby offering a new solution for low-annotation, robust, and deployable intelligent perception. Future work can further explore lightweight architectures, dynamic modality selection, and open-scene generalization.

References

[1] Zong, Yongshuo, Oisin Mac Aodha, and Timothy M. Hospedales. "Self-supervised multimodal learning: A survey." IEEE Transactions on Pattern Analysis and Machine Intelligence 47.7 (2024): 5299-5318.
[2] Chen, Xiaokang, et al. "Context autoencoder for self-supervised representation learning." International Journal of Computer Vision 132.1 (2024): 208-223.
[3] Yang, Zhengeng, et al. "Fully convolutional network-based self-supervised learning for semantic segmentation." IEEE Transactions on Neural Networks and Learning Systems 35.1 (2022): 132-142.
[4] Liu, Yue, et al. "Rethinking self-supervised semantic segmentation: Achieving end-to-end segmentation." IEEE Transactions on Pattern Analysis and Machine Intelligence 46.12 (2024): 10036-10046.
[5] Hoyer, Lukas, et al. "Improving semi-supervised and domain-adaptive semantic segmentation with self-supervised depth estimation." International Journal of Computer Vision 131.8 (2023): 2070-2096.
[6] Zhang, Bowen, et al. "Segvit v2: Exploring efficient and continual semantic segmentation with plain vision transformers." International Journal of Computer Vision 132.4 (2024): 1126-1147.
[7] Wu, Wei, Tao Chu, and Qiong Liu. "Complementarity-aware cross-modal feature fusion network for RGB-T semantic segmentation." Pattern Recognition 131 (2022): 108881.
[8] Wang, Yike, Gongyang Li, and Zhi Liu. "SGFNet: Semantic-guided fusion network for RGB-thermal semantic segmentation." IEEE Transactions on Circuits and Systems for Video Technology 33.12 (2023): 7737-7748.
[9] Lv, Ying, Zhi Liu, and Gongyang Li. "Context-aware interaction network for RGB-T semantic segmentation." IEEE Transactions on Multimedia 26 (2024): 6348-6360.
[10] Peng, Fengguang, et al. "Region-adaptive and context-complementary cross modulation for RGB-T semantic segmentation." Pattern Recognition 147 (2024): 110092.
[11] Liu, Jinfu, et al. "MiLNet: Multiplex interactive learning network for RGB-T semantic segmentation." IEEE Transactions on Image Processing (2025).