Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

HCSC-Net: Hierarchical Contextual Semantic Calibration for CLIP-based Weakly Supervised Semantic Segmentation

Xiaoming Bai

xiaomingbai2025@zua.edu.cn

Zhengzhou University of Aeronautics

Xiaoyan Shao

Zhengzhou University of Aeronautics

Lingling Li

Zhengzhou University of Aeronautics

Xuezhuan Zhao

Zhengzhou University of Aeronautics

Zhenhao Zhao

Zhengzhou University of Aeronautics

Jian Zhang

Zhengzhou University of Aeronautics

Research Article

Keywords: Weakly supervised semantic segmentation, Contrastive Language, Image Pre-training, Class Activation Maps, Vision-Language Learning

Posted Date: July 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10408251/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

HCSC-Net: Hierarchical Contextual Semantic Calibration for CLIP-based Weakly Supervised Semantic Segmentation

Abstract

Weakly supervised semantic segmentation (WSSS) aims to achieve dense pixel-level semantic understanding using only image-level annotations, thereby significantly reducing the dependence on labor-intensive pixel-wise labeling. Recently, CLIP-based WSSS methods have demonstrated promising segmentation performance by leveraging large-scale vision–language pre-training to establish robust cross-modal semantic alignment. However, existing approaches still suffer from insufficient structural consistency, fragmented foreground activation, and severe background interference during dense semantic localization, which often leads to incomplete object regions and unreliable class activation maps (CAMs).To address these limitations, this paper proposes a Hierarchical Contextual Semantic Calibration Network (HCSC-Net) for CLIP-based weakly supervised semantic segmentation. Specifically, a Group-wise Structural Calibration (GSC) module is introduced prior to CLIP visual encoding to enhance low-level structural consistency and suppress noisy spatial responses through group-wise structural modeling and adaptive semantic calibration. Furthermore, a Residual Contextual Semantic Calibration (RCSC) module is designed after dense feature aggregation to progressively refine contextual semantic representations via spatial semantic interaction, channel contextual calibration, and residual contextual modulation. Benefiting from the proposed hierarchical semantic calibration strategy, HCSC-Net effectively improves semantic compactness, regional continuity, and foreground integrity in dense predictions.Extensive experiments conducted on the PASCAL VOC 2012 and MS COCO 2014 benchmarks demonstrate the effectiveness and robustness of the proposed framework. In particular, HCSC-Net achieves 75.8% mIoU on the PASCAL VOC 2012 validation set and 48.1% mIoU on the MS COCO 2014 validation set, outperforming several existing state-of-the-art weakly supervised semantic segmentation methods.

Keywords: Weakly supervised semantic segmentation, Contrastive Language–Image Pre-training, Class Activation Maps,Vision-Language Learning

1 Introduction

Weakly supervised semantic segmentation (WSSS)[1–3] aims to achieve dense, pixel-level semantic understanding using only weak annotations, such as image-level labels, thereby significantly reducing the costly annotation expenses required for fully supervised semantic segmentation. Due to its high annotation efficiency and immense application potential, WSSS has garnered widespread attention in recent years. Existing WSSS methods typically rely on class activation maps (CAMs) to localize distinctive target regions and further generate pseudo-labels for segmentation training. Although significant progress has been made, due to the inherent sparsity of weak supervision, most methods still suffer from issues such as incomplete target activation, inaccurate boundary localization, and severe background interference. Early WSSS methods primarily focused on improving CAM generation through affinity learning, region propagation, or contrastive optimization. For example, affinity-based methods utilize inter-pixel relationships to expand activation regions, while contrastive learning methods aim to enhance intra-class consistency and inter-class discriminability. However, these methods remain fundamentally constrained by the semantic representation capabilities of traditional visual encoders—which are trained on closed datasets. Consequently, the generated activation maps often concentrate solely on the most discriminative parts of objects, leading to semantic fragmentation and instability in dense localization.

In recent years, the development of large-scale vision-language pre-training models has significantly advanced research in WSSS, with methods such as Contrastive Language-Image Pre-training (CLIP) garnering widespread attention. Thanks to its ability to align massive amounts of image-text pairs across modalities, CLIP can establish more robust visual-text semantic associations, thereby effectively improving object localization performance under weakly supervised conditions. Consequently, an increasing number of studies have incorporated CLIP into WSSS tasks. By leveraging the correspondence between image patches and textual semantics, these approaches generate more comprehensive and discriminative CAMs, achieving significantly superior segmentation performance compared to traditional convolutional neural network (CNN)-based frameworks. Existing methods primarily focus on cross-modal semantic alignment and dense feature modeling. MCTformer[4] utilizes Transformers to model long-range dependencies among patch tokens, thereby enhancing the integrity of target regions; WeCLIP[5] improves semantic consistency between local regions and class text through visual-language collaborative learning; CLIMS[6] optimizes the distribution of activated regions through inter-class semantic constraints to mitigate background misactivation issues. Although these methods have achieved significant progress, existing CLIP-based WSSS approaches still suffer from two key limitations: on the one hand, most methods focus excessively on cross-modal semantic alignment while neglecting structural consistency modeling prior to visual encoding, leading to activation fragmentation and spatial response discontinuities in complex backgrounds; on the other hand, after multi-layer feature aggregation, the fused semantic features still lack an effective mechanism for contextual semantic calibration, resulting in incomplete foreground regions, blurred object boundaries, and severe background

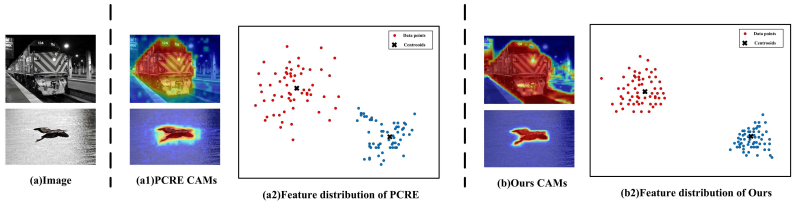


Fig. 1 Comparison of CAM and semantic feature distributions. We use the CLIP-based baseline WSSS method as a reference and visualize the generated CAM and its corresponding semantic feature distributions. In comparison, the proposed HCSC-Net generates more compact feature clusters and more complete foreground activations, demonstrating superior semantic consistency and contextual representation capabilities.

interference. Therefore, how to further enhance structural consistency and contextual semantic modeling capabilities remains a critical research direction for current CLIP-based weakly supervised semantic segmentation.

Existing CLIP-based WSSS methods still tend to generate fragmented activation maps that primarily focus on highly distinctive object regions while neglecting semantically coherent background structures. This lack of semantic calibration further leads to a dispersed feature distribution and weak intra-class compactness in the semantic embedding space, resulting in unstable contextual representations and reduced dense localization quality. In contrast, the proposed Hierarchical Contextual Semantic Calibration Network (HCSC-Ne) effectively enhances structural consistency and contextual semantic compactness through hierarchical semantic calibration. Thanks to this calibration strategy, HCSC-Net generates more complete activation regions and significantly more compact semantic distributions, thereby improving foreground continuity and semantic distinguishability in challenging scenes, as shown in Figure 1.

First, although CLIP-based frameworks improve the semantic alignment between visual patches and textual concepts, they lack effective modeling of structural consistency prior to visual encoding. Since CLIP patch embeddings are highly sensitive to local spatial distributions, noisy textures and background regions can lead to unstable patch representations, resulting in fragmented activation responses and spatial inconsistencies during dense localization. Second, even after multi-level feature aggregation, the resulting dense semantic representations still suffer from insufficient contextual calibration. Existing methods primarily focus on semantic alignment while neglecting contextual semantic compactness and regional continuity, which often results in incomplete object regions and blurred semantic boundaries in complex scenes. To address these issues, we propose a novel HCSC-Net for CLIP-based weakly supervised semantic segmentation. This framework progressively enhances low-level structural consistency and high-level semantic compactness through a hierarchical semantic calibration strategy. Specifically, we introduce a Grouped Structural Calibration (GSC) module prior to visual encoding, which stabilizes local structural distributions and suppresses noisy spatial responses through adaptive group-level region calibration. By enhancing structural consistency before patch embedding, the proposed GSC module effectively improves the reliability of patch-level semantic representations. Furthermore, we designed a Residual Context Semantic Calibration (RCSC) module following dense feature aggregation to optimize contextual semantic representations through

cross-dimensional semantic interaction and residual context modulation. The proposed RCSC module enhances semantic compactness and foreground continuity, thereby generating more complete and accurate dense localization results. Thanks to the hierarchical calibration mechanism, the proposed HCSC-Net effectively mitigates activation fragmentation and semantic ambiguity while improving region continuity and boundary quality. Extensive experiments on the PASCAL VOC 2012 and MS COCO 2014 benchmark datasets demonstrate the superiority of this framework. HCSC-Net achieves a mean IoU (mIoU) of 75.8% on the PASCAL VOC 2012 validation set and 48.1% on the MS COCO 2014 validation set, outperforming various state-of-the-art single-stage WSSS methods based on CLIP. Furthermore, qualitative visualization results further validate that the proposed method can generate more complete object activation regions and clearer semantic boundaries in challenging scenarios.

The main contributions of this work are summarized as follows:

- We propose a novel HCSC-Net for CLIP-based weakly supervised semantic segmentation, which progressively enhances structural consistency and semantic compactness through a hierarchical calibration strategy.
- We introduce a GSC module prior to visual encoding to stabilize patch-level structural representations and suppress noisy spatial responses, thereby improving the consistency of dense localization.
- We design a RCSC module that optimizes the fused dense semantic features through contextual semantic interaction and residual calibration, thereby achieving more complete foreground activation and clearer semantic boundaries.
- Extensive experiments on the PASCAL VOC 2012 and MS COCO 2014 benchmark datasets demonstrate that the proposed HCSC-Net achieves performance comparable to state-of-the-art methods while maintaining an efficient single-stage inference framework.

2 Related Work

2.1 Weakly Supervised Semantic Segmentation

Weakly Supervised Semantic Segmentation (WSSS) aims to achieve pixel-level semantic prediction by utilizing weakly labeled information (such as image-level labels [7–9], bounding boxes[10, 11], doodles[12, 13], and point annotations[14]) to achieve pixel-level semantic predictions, thereby significantly reducing the reliance of fully supervised semantic segmentation on fine-grained pixel-level annotations. Among these, WSSS based on image-level labels has become one of the most extensively studied weakly supervised settings due to its minimal annotation cost and good practical scalability. WSSS methods centered on image-level labels are primarily divided into two major categories: multi-stage and single-stage. Multi-stage methods typically first utilize a classification network to generate CAMs, then employ post-processing strategies such as affinity propagation, region expansion, or pseudo-label optimization to generate high-quality pseudo-labels, and finally train a downstream segmentation network. For example, VPL[15] learns discriminative visual prototypes by introducing

semantic similarity constraints in the visual space and combines them with a region-consistent contrastive strategy, effectively enhancing the semantic completeness and cross-modal alignment capabilities of pseudo-labels; TagCLIP[16], on the other hand, designed two modules—DMAR and CWR—which enhance the robustness of image labels and the quality of pseudo-labels through attention-guided local consistency enhancement and a local-to-global semantic re-identification mechanism, respectively. However, multi-stage methods typically suffer from issues such as complex training workflows and the accumulation of pseudo-label errors during the iterative process, thereby limiting their end-to-end optimization capabilities.

In contrast, single-stage methods directly optimize classification and segmentation tasks jointly within a unified framework, offering advantages such as a simple architecture, efficient training, and consistent inference. For example, DuPL[17] generates diverse pseudo-labels using a dual-student network and combines dynamic thresholds with a Gaussian mixture model to suppress noise interference, thereby effectively mitigating the confirmation bias problem; SeCo[18] proposes a dual-teacher–single-student architecture that enhances semantic discrimination through local–global contrast constraints and category prototype modeling, significantly improving pseudo-label quality and segmentation performance; TagCLIP further utilizes Patch-Level Classification to preserve local spatial information, thereby addressing the issue where traditional CLIP class tokens overlook fine-grained regional features. In recent years, the introduction of contrastive learning and Transformer architectures has further advanced the development of WSSS. Contrastive learning improves semantic consistency by enhancing intra-class compactness and inter-class separability, while Transformers enable more comprehensive object localization through their ability to model long-range dependencies and interact with global context. However, existing methods remain susceptible to background noise and structural inconsistencies in complex scenarios, leading to incomplete target region activation, blurred semantic boundaries, and poor-quality pseudo-labels. In particular, existing single-stage frameworks still lack effective mechanisms for modeling structural consistency and calibrating contextual semantics during dense localization, thereby limiting further improvements in semantic compactness and regional continuity. Therefore, how to further enhance structural modeling capabilities and contextual semantic representation within end-to-end weakly supervised frameworks remains a key research direction in the current WSSS field.

2.2 Vision-Language Pre-training in WSSS

In recent years, advancements in large-scale vision-language pretraining techniques have significantly driven progress in WSSS. Among these, CLIP demonstrates excellent cross-modal semantic alignment capabilities by jointly learning visual and textual representations from massive image-text pairs, and has achieved strong transfer performance across multiple downstream tasks. Building on this, DenseCLIP[19] and MaskCLIP[20] utilize visual–text alignment to achieve dense semantic segmentation; CLIMS[6] enhances visual semantic representation through image–text pairing relationships; CLIP-ES[21] leverages CLIP to generate class gradients to optimize Grad-CAM; and WeCLIP[5] directly employs the CLIP visual encoder to construct a weakly supervised segmentation framework. Compared to traditional CNN-based

methods, existing CLIP-based WSSS methods can generate more comprehensive CAMs through semantic correspondences between patches and text, significantly improving dense localization capabilities under weakly supervised conditions. Some studies further integrate Transformers, multi-layer feature interactions, and token affinity modeling to enhance patch-level semantic consistency and regional continuity. However, most existing methods focus on global image-text semantic alignment while neglecting local structural consistency and contextual semantic refinement. Since CLIP's patch embeddings are sensitive to local spatial distributions, noisy textures in complex backgrounds can easily lead to fragmented activation regions and discontinuous spatial responses. Additionally, after feature fusion, contextual semantic representations still lack effective calibration, resulting in incomplete foreground regions and blurred object boundaries. To address these issues, this paper proposes a hierarchical context-semantic calibration framework HCSC-Net. By performing structural calibration prior to visual encoding and executing context-semantic refinement after feature fusion,HCSC-Net further enhances patch-level consistency, regional semantic continuity, and dense localization capabilities under weakly supervised conditions.

3 Method

3.1 Overview

Figure 2 illustrates the overall architecture of the proposed HCSC-Net, which follows the CLIP-based single-stage weakly supervised semantic segmentation paradigm. The proposed framework primarily integrates two hierarchical semantic calibration modules: GSC module and the RCSC module, to progressively enhance structural consistency and contextual semantic compactness during the dense localization process. By performing structural refinement prior to visual encoding and executing contextual semantic calibration after dense feature aggregation, HCSC-Net is capable of directly generating high-quality semantic representations and reliable pseudo-labels in an end-to-end manner, thereby effectively alleviating the issues of activation region fragmentation and semantic ambiguity that are prevalent in existing CLIP-based WSSS methods. The overall training process primarily consists of the following stages.

- **Structural calibration.** The proposed GSC module performs group-wise structural modeling before visual encoding by dividing feature maps into multiple groups for adaptive calibration. This process suppresses noisy responses and enhances local structural consistency, thereby improving the reliability of patch-level semantic representations.
- **Visual-to-language semantic extraction.** The calibrated visual features are fed into a frozen CLIP visual encoder, while category labels are transformed into text prompts and processed by the CLIP text encoder. Semantic similarity between visual and textual features is subsequently computed to generate initial semantic localization maps.

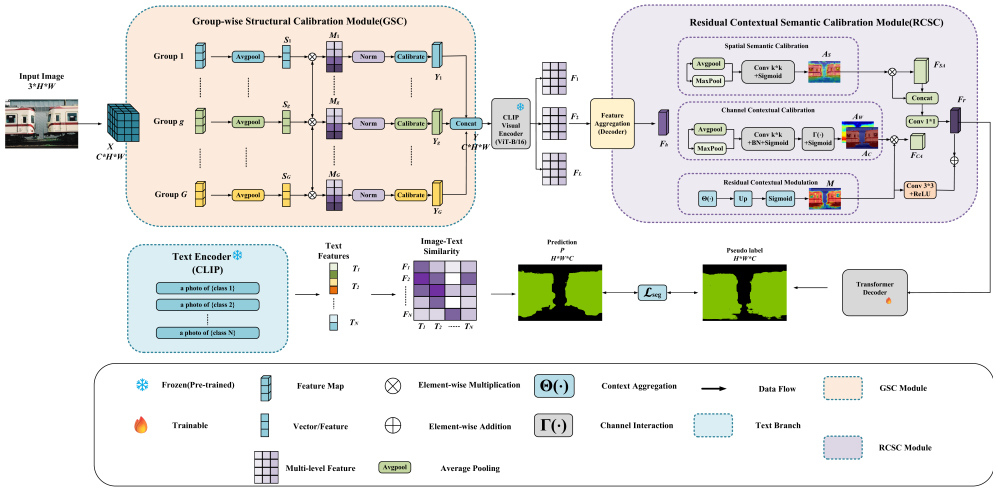


Fig. 2 The overall architecture of the proposed framework, which mainly consists of the GSC module and the RCSC module. These two modules collaboratively enhance feature representation and semantic localization under complex weakly supervised scenarios. Given an input image, multi-level visual features are first extracted by a frozen CLIP visual encoder. The GSC module divides feature maps into multiple groups to model diverse structural semantics, where each group performs average pooling, normalization, and adaptive calibration to capture complementary structural information and suppress noisy responses. The calibrated group-wise features are then concatenated to obtain enhanced visual representations. Meanwhile, text embeddings generated by the CLIP text encoder are utilized to establish cross-modal semantic alignment, thereby producing the initial semantic localization map for dense prediction.

- **Dense feature aggregation.** Multi-level visual features are fused through a feature aggregation decoder to obtain dense semantic representations with richer contextual information and enhanced regional continuity.
- **Residual contextual semantic calibration.** The proposed RCSC module progressively refines dense semantic representations through spatial semantic calibration, channel contextual interaction, and residual contextual modulation, thereby improving semantic compactness, foreground continuity, and boundary integrity.
- **Pseudo-label generation and segmentation learning.** The refined semantic predictions are converted into reliable pseudo-labels to supervise the Transformer decoder. The proposed framework directly performs end-to-end dense segmentation optimization under weak supervision, effectively avoiding cumulative errors caused by multi-stage iterative refinement.

By jointly incorporating hierarchical structure calibration and contextual semantic refinement mechanisms, HCSC-Net effectively enhances patch-level consistency, semantic compactness, and dense localization quality under weakly supervised conditions. The proposed framework generates more complete region-of-interest activations and more discriminative semantic representations, thereby achieving superior semantic segmentation performance in complex scenes. During training, the entire framework undergoes end-to-end joint optimization based on the loss function defined in Section 3.4.

3.2 Group-wise Structural Calibration module

To alleviate the structural inconsistency and insufficient spatial discrimination caused by complex backgrounds and noisy semantic responses, we propose a lightweight yet effective GSC module to enhance the structural representation capability of visual features before CLIP visual encoding. As illustrated in Fig 2, the proposed GSC module performs group-wise structural modeling and adaptive spatial calibration, enabling the network to capture complementary structural semantics from different channel groups while suppressing irrelevant background interference.

Given the input feature map $X \in \mathbb{R}^{C\times H\times W}$,the feature channels are first evenly divided into G groups along the channel dimension:

$$X = \{X_1, X_2, ..., X_G\}, X_g \in \mathbb{R}^{\frac{C}{G}\times H\times W} \quad (1)$$

where each grouped feature focuses on distinct structural characteristics and regional semantic responses, thereby reducing channel redundancy and improving representation diversity.

For each group, global structural statistics are first extracted through adaptive average pooling and subsequently utilized to recalibrate the grouped feature representation:

$$\hat{X}_g = X_g \odot \mathrm{AvgPool}(X_g) \quad (2)$$

where $\odot$ denotes element-wise multiplication. This operation strengthens structurally informative activations while suppressing noisy background responses.

Subsequently, channel-wise aggregation is performed to generate the spatial structural response map:

$$M_g = \sum_{c=1}^{C/G} \hat{X}_g^{(c)} \quad (3)$$

where $M_g \in \mathbb{R}^{1\times H\times W}$ represents the spatial structural distribution of the current group.

To alleviate activation inconsistency across different spatial regions, the response map is further normalized and adaptively recalibrated as:

$$W_g = \sigma\left(\alpha_g \frac{M_g - \mu(M_g)}{\sigma(M_g) + \varepsilon} + \beta_g\right) \quad (4)$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation operations, respectively, ε is a small constant for numerical stability, α_g and β_g are learnable scaling and bias parameters corresponding to the g -th group.

The obtained structural attention weights are then applied to the original grouped feature to generate the calibrated representation:

$$Y_g = X_g \odot W_g \quad (5)$$

Algorithm 1 Algorithm 1 Algorithm1 GSC Module Detailed Structure

1.Input: Feature map $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$
2.Step 1: Group-wise Structural Decomposition
3. Divide the input feature map into G groups along the channel dimension:
4. $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_G\}, \quad \mathbf{X}_g \in \mathbb{R}^{\frac{C}{G} \times H \times W}$
5.Step 2: Structural Feature Calibration
6. Extract global structural statistics via adaptive average pooling:
7. $\mathbf{S}_g = \text{AvgPool}(\mathbf{X}_g)$
8. Perform structural feature recalibration using element-wise multiplication:
9. $\hat{\mathbf{X}}_g = \mathbf{X}_g \odot \mathbf{S}_g$
10.Step 3: Spatial Structural Response Generation
11. Aggregate channel-wise structural responses:
12. $\mathbf{M}_g = \sum_{c=1}^{C/G} \hat{\mathbf{X}}_g^{(c)}$
13. Normalize and adaptively recalibrate spatial responses:
14. $\mathbf{W}_g = \sigma\left(\alpha_g \frac{\mathbf{M}_g - \mu(\mathbf{M}_g)}{\sigma(\mathbf{M}_g) + \varepsilon} + \beta_g\right)$
15.Step 4: Group-wise Structural Enhancement
16. Generate calibrated structural representation:
17. $\mathbf{Y}_g = \mathbf{X}_g \odot \mathbf{W}_g$
18.Step 5: Output Aggregation
19. Concatenate all calibrated group features:
20. $\mathbf{Y} = \text{Concat}(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_G)$
21.Output: Structurally calibrated feature representation $\mathbf{Y}$

Finally, the outputs from all groups are concatenated along the channel dimension to obtain the final structurally calibrated feature representation:

$$Y = \text{Concat}(Y_1, Y_2, ..., Y_G) \quad (6)$$

Different from conventional channel attention mechanisms that mainly emphasize inter-channel dependency modeling, the proposed GSC module explicitly explores group-wise structural distributions and spatial semantic variations. By introducing grouped structural normalization and adaptive calibration, the proposed module effectively enhances structural consistency and suppresses noisy background interference before patch embedding. Such a design is particularly beneficial for weakly supervised semantic segmentation, where dense localization quality heavily depends on the structural reliability of patch-level representations.

Furthermore, owing to its lightweight architecture and parameter-efficient design, the proposed GSC module can be seamlessly integrated into the frozen CLIP visual encoder without introducing significant computational overhead. Experimental results demonstrate that the proposed GSC module effectively improves activation completeness, regional continuity, and boundary quality, thereby generating more reliable semantic localization maps under complex scene conditions.

3.3 Residual Contextual Semantic Calibration module

Although dense semantic representations generated after feature aggregation can provide rich contextual information, they may still suffer from semantic ambiguity, incomplete foreground activation, and blurred object boundaries under complex scene conditions. To further improve contextual semantic compactness and regional continuity, we propose a RCSC module, as illustrated in Fig. 2. The proposed module jointly exploits spatial semantic calibration, channel contextual interaction, and residual contextual modulation to progressively refine dense semantic representations and suppress background interference.

Given the aggregated feature representation $F \in \mathbb{R}^{C\times H\times W}$, an initial feature transformation is first performed using a convolution operation followed by nonlinear activation:

$$F_h = \delta(\mathrm{Conv}_{3\times 3}(F)) \quad (7)$$

where $\delta(\cdot)$ denotes the LeakyReLU activation function. The transformed feature F_h serves as the shared semantic representation for subsequent calibration branches.

Spatial Semantic Calibration To enhance spatial semantic localization capability and improve foreground–background discrimination, the proposed RCSC module first introduces a spatial semantic calibration branch. Specifically, channel-wise average pooling and max pooling are simultaneously employed to aggregate complementary spatial contextual statistics, and the generated descriptors are concatenated to produce the

$$A_s = \sigma\left(\mathrm{Conv}_{k\times k}\left(\mathrm{Concat}(\mathrm{AvgPool}_c(F_h), \mathrm{MaxPool}_c(F_h))\right)\right) \quad (8)$$

where $\sigma(\cdot)$ denotes the sigmoid activation function. The spatially calibrated semantic representation is then obtained by:

$$F_{SA} = F_h \odot A_s \quad (9)$$

Through adaptive spatial semantic weighting, the proposed branch effectively emphasizes informative foreground regions while suppressing noisy background activations, thereby improving semantic localization completeness and spatial continuity.

Channel Contextual Calibration Besides spatial semantic modeling, contextual semantic interaction across feature channels also plays an essential role in dense semantic representation learning. Therefore, a channel contextual calibration branch is further introduced to model inter-channel semantic dependency and improve semantic discriminability. Different from conventional global channel attention mechanisms, the proposed module jointly performs adaptive average pooling and adaptive max pooling along the spatial width dimension. The pooled contextual features are concatenated and transformed through convolution and normalization operations:

$$F_c = \delta\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times 1}(\mathrm{Concat}(\mathrm{AvgPool}_w(F_h), \mathrm{MaxPool}_w(F_h)))\right)\right) \quad (10)$$

where $\mathrm{BN}(\cdot)$ denotes Batch Normalization. Subsequently, the transformed contextual feature is decomposed into two channel-aware excitation branches:

$$A_h = \sigma(F_h(F_c)), A_w = \sigma(F_w(F_c)) \quad (11)$$

where $\mathcal{F}_h(\cdot)$ and $\mathcal{F}_w(\cdot)$ denote lightweight channel excitation operations implemented using bottleneck convolution layers. The final channel-context calibrated representation is obtained as:

$$F_{CA} = F_h \odot A_h \odot A_w \quad (12)$$

By jointly modeling channel contextual dependency and semantic interaction, the proposed branch dynamically enhances discriminative semantic channels while suppressing irrelevant contextual responses.

Residual Contextual Modulation To further improve semantic compactness and contextual continuity, the outputs from the spatial semantic calibration branch and channel contextual calibration branch are first concatenated and compressed through convolution operations:

$$F_r = \mathrm{Conv}_{3\times3}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{Concat}(F_{SA}, F_{CA})\right)\right) \quad (13)$$

Meanwhile, a residual contextual modulation branch is introduced to capture large-range contextual dependency. Specifically, average pooling is first employed to enlarge the receptive field and aggregate broader semantic context:

$$F_{ctx} = \mathrm{BN}\left(\mathrm{Conv}_{3\times3}\left(\mathrm{AvgPool}(F_h)\right)\right) \quad (14)$$

The contextual feature is subsequently upsampled to the original spatial resolution and fused with the current semantic representation:

$$M = \sigma(F_h + \cup\mathrm{psample}(F_{ctx})) \quad (15)$$

The generated contextual modulation map M is then utilized to recalibrate the fused semantic representation:

$$F_o = F_r \odot M \quad (16)$$

Finally, a residual connection is introduced to preserve the original semantic structure and stabilize feature propagation during optimization:

$$F_{out} = \delta(\mathrm{Conv}_{3\times3}(F_o) + F) \quad (17)$$

Compared with conventional attention-based refinement strategies, the proposed RCSC module simultaneously models spatial semantic dependency, channel contextual interaction, and large-range contextual modulation within a unified residual semantic calibration framework. Such a hierarchical calibration mechanism effectively improves semantic compactness, foreground continuity, and boundary integrity while suppressing cluttered background interference.

Furthermore, owing to its lightweight residual architecture and efficient contextual modeling capability, the proposed RCSC module can be seamlessly integrated into the dense feature aggregation stage without introducing significant computational

overhead. Experimental results demonstrate that the proposed module substantially improves semantic localization quality and generates more complete and discriminative foreground activation regions under weak supervision.

Algorithm 2 Algorithm 2 Algorithm1 RCSC Module Detailed Structure

1.Input: Aggregated feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$
2.Step 1: Initial Semantic Embedding
3. Generate shared semantic representation:
4. $\mathbf{F}_h = \delta(\mathrm{Conv}_{3\times3}(\mathbf{F}))$
5.Step 2: Spatial Semantic Calibration
6. Aggregate spatial contextual descriptors:
7. $\mathbf{A}_s = \sigma\left(\mathrm{Conv}_{k\times k}\left(\mathrm{Concat}(\mathrm{AvgPool}_c(\mathbf{F}_h), \mathrm{MaxPool}_c(\mathbf{F}_h))\right)\right)$
8. Generate spatially calibrated semantic feature:
9. $\mathbf{F}_{SA} = \mathbf{F}_h \odot \mathbf{A}_s$
10.Step 3: Channel Contextual Calibration
11. Extract contextual semantic statistics:
12. $\mathbf{F}_c = \delta\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{Concat}(\mathrm{AvgPool}_w(\mathbf{F}_h), \mathrm{MaxPool}_w(\mathbf{F}_h))\right)\right)\right)$
13. Generate channel-aware excitation weights:
14. $\mathbf{A}_h = \sigma(\mathcal{F}_h(\mathbf{F}_c)), \quad \mathbf{A}_w = \sigma(\mathcal{F}_w(\mathbf{F}_c))$
15. Obtain channel-context calibrated representation:
16. $\mathbf{F}_{CA} = \mathbf{F}_h \odot \mathbf{A}_h \odot \mathbf{A}_w$
17.Step 4: Residual Contextual Modulation
18. Fuse spatial and channel semantic features:
19. $\mathbf{F}_r = \mathrm{Conv}_{3\times3}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{Concat}(\mathbf{F}_{SA}, \mathbf{F}_{CA})\right)\right)$
20. Generate contextual modulation map and refine semantic feature:
21. $\mathbf{M} = \sigma\left(\mathbf{F}_h + \mathrm{Upsample}\left(\mathrm{BN}\left(\mathrm{Conv}_{3\times3}(\mathrm{AvgPool}(\mathbf{F}_h))\right)\right)\right)$
22. $\mathbf{F}_o = \mathbf{F}_r \odot \mathbf{M}$
23.Step 5: Residual Semantic Refinement
24. Preserve semantic continuity via residual learning:
25. $\mathbf{F}_{out} = \delta(\mathrm{Conv}_{3\times3}(\mathbf{F}_o) + \mathbf{F})$
26.Output: Refined semantic feature representation $\mathbf{F}_{out}$

3.4 Loss Function

To effectively optimize the proposed HCSC-Net under weak supervision, we adopt a hybrid optimization objective that jointly constrains semantic prediction accuracy and contextual feature consistency. As illustrated in Fig 2, the overall training objective mainly consists of a segmentation supervision loss and an affinity consistency regularization loss. The segmentation loss supervises dense semantic prediction generated by the Transformer decoder, while the affinity regularization term constrains contextual semantic consistency between pixel-level feature representations. The overall

optimization objective can be formulated as:

$$L_{total} = L_{seg} + \lambda L_{aff} \quad (18)$$

where L_{seg} denotes the segmentation supervision loss,L_{aff} represents the affinity consistency loss, and λ is a balancing coefficient controlling the contribution of affinity regularization.

Segmentation Supervision Loss Following the single-stage weakly supervised semantic segmentation paradigm, the proposed framework first generates pseudo labels from the refined semantic localization maps produced by the GSC and RCSC modules. These pseudo labels are subsequently utilized to supervise the Transformer decoder for dense semantic prediction.

Given the predicted segmentation probability map $P \in \mathbb{R}^{C\times H\times W}$ and the generated pseudo label Y, the segmentation supervision objective is formulated using the pixel-wise cross-entropy loss:

$$L_{seg} = -\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{c} Y_{i,j}^{c}\log\left(P_{i,j}^{c}\right) \quad (19)$$

where C denotes the number of semantic categories, and $P_{i,j}^{c}$ represents the predicted probability of pixel (i, j) belonging to category c.

The segmentation supervision loss encourages the network to generate semantically accurate dense predictions while progressively improving object localization quality under image-level supervision.

Affinity Consistency Loss Although pseudo labels generated from semantic localization maps provide effective supervision signals, weakly supervised semantic segmentation still suffers from semantic fragmentation and inconsistent contextual responses under complex scenes. To alleviate this issue, the proposed framework further introduces an affinity consistency loss to constrain semantic correlation between dense feature representations.

Specifically, let $F_{att} \in \mathbb{R}^{C\times H\times W}$ denote the refined contextual semantic feature generated by the RCSC module. The feature representation is first reshaped and normalized into a pixel-level semantic embedding:

$$\hat{F}_{flat} = \frac{\text{Reshape}(F_{att})}{\parallel \text{Reshape}(F_{att}) \parallel_2}, \hat{F}_{flat} \in \mathbb{R}^{C\times HW} \quad (20)$$

The normalized feature representations are subsequently utilized to compute the pixel-wise semantic affinity matrix:

$$A = \hat{F}_{flat}^{T}\hat{F}_{flat}, A \in \mathbb{R}^{HW\times HW} \quad (21)$$

where A represents the semantic affinity relationship between different spatial locations. To stabilize optimization and suppress noisy semantic responses, the affinity

prediction is further normalized and activated through a sigmoid function:

$$A_{pred} = \sigma\left(\alpha(A - \mu(A))\right) \quad (22)$$

where $\mu(\cdot)$ denotes the mean operation,$\sigma(\cdot)$ represents the sigmoid activation function, and α is a scaling coefficient controlling affinity distribution sharpness.

The affinity consistency loss is then formulated as:

$$L_{aff} = \frac{1}{N} \parallel A_{pred} - A_{gt} \parallel_2^2 \quad (23)$$

where A_{gt} denotes the affinity supervision generated from pseudo labels, and N represents the total number of affinity elements.

The proposed affinity consistency regularization encourages pixels belonging to the same semantic category to maintain similar contextual representations while enlarging semantic discrepancy between different categories. Consequently, the network can effectively enhance regional semantic compactness and suppress fragmented activation responses.

Overall Optimization Objective By jointly optimizing segmentation prediction and affinity consistency, the proposed hybrid objective simultaneously enhances semantic localization accuracy and contextual feature consistency. Benefiting from the collaborative optimization of the GSC module, RCSC module, and affinity consistency supervision, the proposed HCSC-Net effectively improves foreground continuity, semantic compactness, and boundary integrity under complex scene conditions, thereby generating more reliable dense semantic localization results.

4 Experiments and Results

4.1 Experimental Settings

Datasets and Metrics. Following the majority of existing weakly supervised semantic segmentation methods based on image-level labels[22, 23], we evaluate the proposed improved method on two standard benchmark datasets: PASCAL VOC 2012[24] and MS COCO 2014[25]. For PASCAL VOC 2012, this paper combines the official training set with the additional SBD extension[26], comprising a total of 10,582 training images, 1,449 validation images, and 1,456 test images, covering 20 ground-truth classes. MS COCO 2014 comprises approximately 82,081 training images and 40,137 validation images, covering 80 semantic categories. Mean Intersection over Union (mIoU) is used as the primary evaluation metric, and performance results are reported separately for the validation and test sets.

Implementation Details. Throughout this study, the VIT-B visual backbone network[27, 28]was used as the image encoder. This paper extracts multiscale features from four Transformer blocks (each layer has a feature dimension of 768) and maps them to 256-dimensional embeddings via a learned projection layer for subsequent decoder processing. During training on the PASCAL VOC 2012 dataset, the batch size was set to 4, with a maximum of 30,000 iterations. On the MS COCO-2014 dataset,

the batch size was set to 8, with a maximum of 80,000 iterations. During training, input images are first randomly scaled by a factor between [0.5, 2.0], then cropped to a size of 320×320. Additionally, the long edge of the image is scaled to a random value within the range [512, 2048], and all semantically ambiguous pixels are assigned the ignore label 255.

4.2 Comparison with State-of-the-art Methods

Table 1 presents a quantitative comparison between the proposed HCSC-Net and the state-of-the-art WSSS methods on the PASCAL VOC 2012 and MS COCO 2014 benchmark datasets. The experimental results demonstrate that, under image-level supervision, the proposed method achieves excellent segmentation performance, validating the effectiveness of the proposed structure-context semantic calibration strategy. On the PASCAL VOC 2012 dataset, the proposed HCSC-Net achieves a mIoU of 75.8% on the validation set and 77.8% on the test set, outperforming most existing single-stage and multi-stage WSSS methods. Compared to a strong CLIP-based baseline model (which achieved mIoU of 75.5% and 75.9% on the validation and test sets, respectively), HCSC-Net improved segmentation accuracy by 0.3% and 1.9%, respectively. This performance improvement demonstrates that the proposed GSC and RCSC modules effectively enhance semantic localization capabilities and boundary integrity under weakly supervised conditions.

Furthermore, HCSC-Net outperforms several recent state-of-the-art methods across all metrics. Compared to the dense image-text alignment method DIAL[29], our proposed method improves mIoU from 74.5% to 75.8% on the validation set and from 74.9% to 77.8% on the test set, representing increases of 1.3% and 2.9%, respectively. Compared to the multi-stage method PSDPM, HCSC-Net achieves a 1.7% improvement on the validation set and a 2.9% improvement on the test set. Similarly, compared to CPAL, which incorporates context-aware prototype learning to enhance semantic understanding, the proposed framework still achieves significant improvements of 1.3% and 3.1% on the validation and test sets, respectively. These results demonstrate that the proposed framework is more effective in modeling structural semantics, suppressing background interference, and optimizing object boundaries. Furthermore, the proposed method demonstrates strong generalization capabilities on the more challenging MS COCO 2014 dataset. HCSC-Net achieves a mIoU of 48.1% on the validation set, outperforming WeCLIP (46.4%), CPAL (46.8%), and DIAL (44.4%) by 1.7%, 1.3%, and 3.7%, respectively. Compared to recent competitive methods such as SeCo and PSDPM, the proposed framework maintains excellent performance even in complex multi-object scenes and under sparse semantic distributions. This improvement validates that HCSC-Net can effectively capture long-range contextual dependencies and complex semantic relationships, thereby demonstrating stronger robustness and generalization capabilities in large-scale red-out semantic segmentation tasks.

As shown in Figure 3, visualization results on the PASCAL VOC 2012 and MS COCO 2014 validation sets further validate the effectiveness and robustness of the proposed HCSC-Net. Compared to the state-of-the-art method PCRE, our approach generates more complete, continuous, and semantically consistent segmentation results in complex scenes. Thanks to the synergistic interaction between the GSC and RCSC

Table 1 Segmentation comparisons on VOC and COCO. Sup. denotes the supervision type. Net. denotes the backbone network. $\mathcal{I}$: Image-level labels; $\mathcal{SA}$: Saliency maps; $\mathcal{L}$: Language.

Methods	Sup.	Net.	VOC		COCO
			val (%)	test (%)	val (%)
Multi-stage WSSS methods					
L2G [30] CVPR'2022	$\mathcal{I}+\mathcal{SA}$	RN101	72.1	71.7	44.2
RCA [31] CVPR'2023	$\mathcal{I}+\mathcal{SA}$	RN38	72.2	72.8	36.8
OCR [32] CVPR'2023	$\mathcal{I}$	RN38	72.7	72.0	42.5
BECO [33] CVPR'2023	$\mathcal{I}$	RN101	73.7	73.5	45.1
CTI [34] CVPR'2024	$\mathcal{I}$	RN101	74.1	73.2	45.4
CLIMS [6] CVPR'2022	$\mathcal{I}+\mathcal{L}$	RN101	70.4	70.0	–
CLIP-ES [21] CVPR'2023	$\mathcal{I}+\mathcal{L}$	RN101	72.2	72.8	45.4
PSDPM [35] CVPR'2024	$\mathcal{I}+\mathcal{L}$	RN101	74.1	74.9	47.2
CPAL [36] CVPR'2024	$\mathcal{I}+\mathcal{L}$	RN101	74.5	74.7	46.8
Single-stage WSSS methods					
AFA [37] CVPR'2022	$\mathcal{I}$	MiT-B1	66.0	66.3	38.9
ViT-PCM [38] ECCV'2022	$\mathcal{I}$	ViT-B	70.3	70.9	–
ToCo [39] CVPR'2023	$\mathcal{I}$	ViT-B	71.1	72.2	42.3
DuPL [17] CVPR'2024	$\mathcal{I}$	ViT-B	73.3	72.8	44.6
SeCo [18] CVPR'2024	$\mathcal{I}$	ViT-B	74.0	73.8	46.7
DIAL [29] ECCV'2024	$\mathcal{I}+\mathcal{L}$	ViT-B	74.5	74.9	44.4
WeCLIP [5] CVPR'2024	$\mathcal{I}+\mathcal{L}$	ViT-B	74.9	75.2	46.4
PCRE [40] CVPR'2025	$\mathcal{I}$	ViT-B	75.5	75.9	47.2
HCSC-Net (w/o CRF)	$\mathcal{I}+\mathcal{L}$	**ViT-B**	**75.8**	**77.8**	**48.1**

modules, HCSC-Net demonstrates stronger object localization and context modeling capabilities in multi-object, complex background, and fine-grained structural scenarios, effectively preserving object boundaries and local details. The results demonstrate that for categories such as trains, animals, human bodies, and complex indoor scenes, our method generates clearer and more coherent semantic regions, whereas the baseline methods are prone to issues such as missing objects, blurred boundaries, and background misclassification. Furthermore, in complex scenes on the COCO dataset, HCSC-Net remains capable of effectively suppressing background noise and achieving more natural and smooth segmentation boundaries for small objects and elongated structural regions. Overall, the visualization results fully demonstrate the advantages of the proposed structure–context joint calibration strategy in terms of semantic completeness, contextual consistency, and boundary refinement.

Table 2 presents the quantitative comparison results of the quality of the original CAM seeds on the PASCAL VOC 2012 training set. Experimental results demonstrate that the proposed HCSC-Net exhibits superior semantic localization capabilities among existing weakly supervised semantic segmentation methods. Specifically, without introducing additional CRF post-processing, HCSC-Net achieves an mIoU of 77.0%, outperforming current mainstream multi-stage and single-stage methods. Compared to WeCLIP (75.4%), a representative CLIP-based method, our method achieves a 1.6% improvement and also outperforms recent representative methods such as DIAL (75.2%), SeCo (74.8%), DuPL (75.0%), and PCRE (76.7%). These performance improvements fully validate the effectiveness of the proposed structure-context

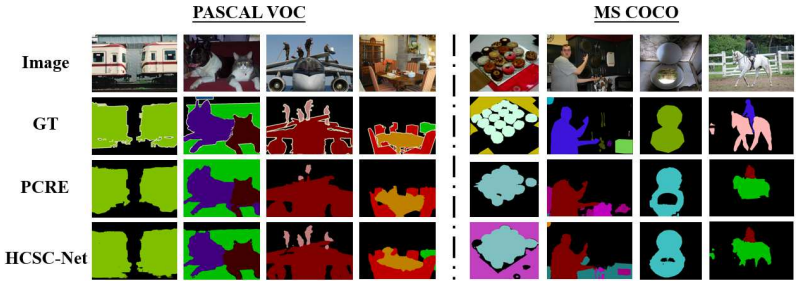


Fig. 3 Our method was qualitatively compared with WeCLIP [44] on the PASCAL VOC 2012 and MS COCO-2014 datasets. Our method generates more detailed visualizations.

semantic calibration strategy in enhancing dense semantic localization capabilities. Specifically, the GSC module effectively improves feature structural consistency and suppresses background interference by modeling complementary semantic information among different feature groups; meanwhile, the RCSC module further enhances contextual semantic alignment and target boundary integrity through a combination of spatial-channel attention modeling and a residual context refinement mechanism. Thanks to the synergistic optimization of these two modules, HCSC-Net is able to generate activation regions that are more complete, accurate, and semantically consistent, thereby providing higher-quality pseudo-labels for subsequent weakly supervised segmentation learning.

Table 2 CAM seed comparisons on the VOC train set. $\mathcal{M}$: multi-stage methods; $\mathcal{S}$: single-stage methods. †: our reproduction following official codes. HCSC-Net *: HCSC-Net in a training-free manner.

Methods	Type	Sup.	Net.	VOC Train (%)
Training-required WSSS methods				
ReCAM [41] CVPR'22	$\mathcal{M}$	$\mathcal{I}$	RN101	54.8
FPR [1] CVPR'23	$\mathcal{M}$	$\mathcal{I}$	RN101	63.8
LPCAM [42] CVPR'23	$\mathcal{M}$	$\mathcal{I}$	RN50	65.3
MCTformer+ [4] CVPR'23	$\mathcal{M}$	$\mathcal{I}$	RN38	68.8
SFC [43] CVPR'24	$\mathcal{M}$	$\mathcal{I}$	RN101	64.7
CTI [34] CVPR'22	$\mathcal{M}$	$\mathcal{I}$	RN101	69.5
AFA [37] CVPR'23	$\mathcal{S}$	$\mathcal{I}$	MiT-B1	65.0
ViT-PCM [38] CVPR'24	$\mathcal{S}$	$\mathcal{I}$	ViT-B	67.7
† ToCo [39] CVPR'24	$\mathcal{S}$	$\mathcal{I}$	ViT-B	71.6
† DuPL [17] CVPR'22	$\mathcal{S}$	$\mathcal{I}$	ViT-B	75.0
SeCo [18] CVPR'24	$\mathcal{S}$	$\mathcal{I}$	ViT-B	74.8
CLIMS [6] ECCV'22	$\mathcal{M}$	$\mathcal{I}+\mathcal{L}$	RN101	56.6
POLE [44] CVPR'23	$\mathcal{M}$	$\mathcal{I}+\mathcal{L}$	RN50	59.0
CPAL [36] CVPR'24	$\mathcal{M}$	$\mathcal{I}+\mathcal{L}$	RN101	71.9
DIAL [29] ECCV'24	$\mathcal{S}$	$\mathcal{I}+\mathcal{L}$	ViT-B	75.2
† WeCLIP [5] CVPR'24	$\mathcal{S}$	$\mathcal{I}+\mathcal{L}$	ViT-B	75.4
† PCRE [40] CVPR'25	$\mathcal{S}$	$\mathcal{I}$	ViT-B	77.6
HCSC-Net (w/o CRF)	$\mathcal{S}$	$\mathcal{I}+\mathcal{L}$	**ViT-B**	**77.0**

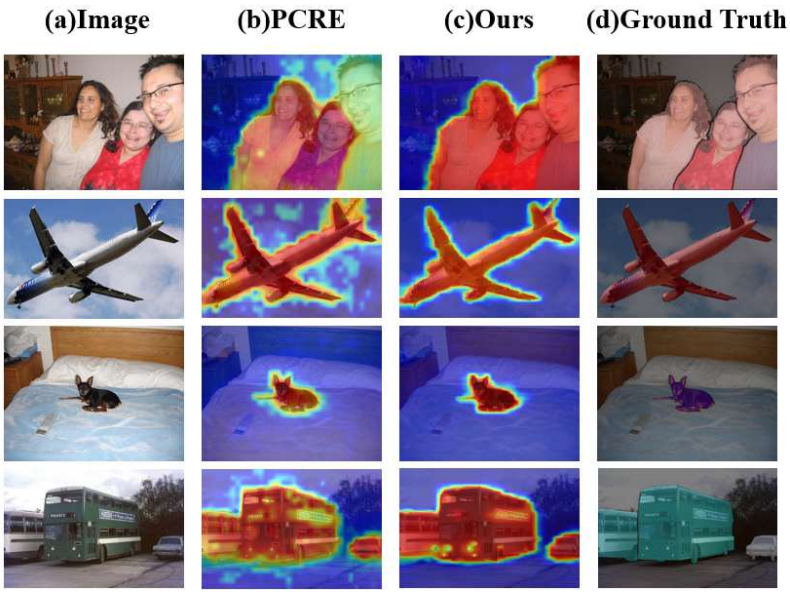


Fig. 4 Comparison of CAM performance between PCRE and HCSC-Net. HCSC-Net effectively suppresses over-activation while achieving more comprehensive and accurate localization of the target region.

Figure 4 shows the qualitative comparison results of CAMs on the PASCAL VOC 2012 training set. Compared to PCRE, our method generates more complete, continuous, and consistent activation responses with the target regions in scenarios involving multiple instances, large-scale objects, small objects, and complex backgrounds. PCRE tends to focus only on the most discriminative local regions, which can lead to activation fragmentation and false background responses. In contrast, our method more accurately covers the entire regions of targets such as human figures, airplanes, birds, and buses, while preserving clearer boundary information and effectively suppressing background interference. Consequently, the CAMs generated by our method exhibit higher consistency with ground-truth annotations and demonstrate superior performance in terms of region completeness and boundary fidelity. The above results demonstrate that our method effectively mitigates the issue of incomplete activations prevalent in traditional CAM-based WSSS and provides higher-quality pseudo-labels for subsequent segmentation models.

4.3 Ablation Studies

Table 3 presents the ablation experiment results of the proposed HCSC-Net on the PASCAL VOC 2012 validation set, which are used to validate the effectiveness of the GSC module and the RCSC module. The experimental results show that when using only the frozen CLIP backbone network as the baseline model, the model achieved an mIoU of 74.3%, with Precision and Recall at 84.0% and 85.7%, respectively. After introducing the GSC module, model performance improved significantly, with mIoU

Table 3 Ablation study of HCSC-Net on the PASCAL VOC 2012 val set.

Conditions	GSC	RCSC	Precision (%)	Recall (%)	mIoU (%)
Baseline	–	–	84.0	85.7	74.3
w/GSC	√	–	85.4	88.2	75.3
w/RCSC	–	√	85.6	87.8	75.4
HCSC-Net	√	√	**85.5**	**88.5**	**75.8**

increasing to 75.3% and Precision and Recall rising to 85.4% and 87.1%, respectively, indicating that this module effectively enhances structural semantic expression capabilities and improves the structural integrity and localization accuracy of target regions. Similarly, when only the RCSC module was introduced, the model achieved an mIoU of 75.4%, with Precision and Recall reaching 85.6% and 87.8%, respectively, validating the effectiveness of the residual context semantic calibration mechanism in enhancing contextual consistency, suppressing background noise, and optimizing semantic discrimination capabilities. When the GSC and RCSC modules are combined, the complete HCSC-Net achieves the best performance, reaching 85.5%, 88.5%, and 75.8% for Precision, Recall, and mIoU, respectively. Compared to the baseline model, the complete framework achieved a 1.5% improvement in mIoU, while also demonstrating further gains over models using only a single module. This fully demonstrates that GSC and RCSC are highly complementary in structural feature modeling and contextual semantic calibration, and can work together to enhance object localization, boundary refinement, and overall semantic consistency in weakly supervised semantic segmentation.

Table 4 illustrates the impact of different clustering settings on model performance within the GSC module, to further analyze the effectiveness of clustering strategies in structural semantic modeling. The experimental results indicate that a reasonable number of clusters can effectively enhance the model's semantic representation capabilities and object localization accuracy. Without group structure calibration, the baseline model achieved an mIoU of 74.3%, a Precision of 84.0%, and a Recall of 85.7% on the PASCAL VOC 2012 validation set. After introducing the GSC module, all group sizes resulted in performance improvements. Specifically, when the number of groups was set to 8, the model achieved optimal performance, with Precision, Recall, and mIoU reaching 85.4%, 88.2%, and 75.3%, respectively. Compared to the baseline model, mIoU improved by 1.0%, while Recall increased by 2.5%. This indicates that an appropriate grouping strategy can effectively enhance the structural semantic complementarity between different channels and improve the integrity and contextual consistency of the target regions. When the number of groups is small (e.g., 2 groups), the model struggles to fully capture complex semantic structures, resulting in limited performance gains; conversely, when the number of clusters is too high (e.g., 16 clusters), features are overly segmented, leading to weakened local semantic relationships and consequently causing a performance decline. Overall, the experimental results indicate that a moderate number of clusters achieves a better balance between

Table 4 Number of groups in GSC on the PASCAL VOC 2012 val set.

Group	Baseline	#2	#4	#8	#16
Precision (%)	84.0	85.5	85.2	**85.4**	85.4
Recall (%)	85.7	86.4	87.3	**88.2**	87.1
mIoU (%)	74.3	74.0	75.2	**75.3**	75.2

Table 5 Attention path design on the PASCAL VOC 2012 val set.

Conditions	SA	CA	RCM	Precision (%)	Recall (%)	mIoU (%)
Baseline	–	–	–	84.0	85.7	74.3
+SA	√	–	–	85.3	87.7	75.2
+SA+CA	√	√	–	85.5	87.7	75.2
RCSC	√	√	√	**85.6**	**87.8**	**75.4**

structural semantic representation and feature consistency, further validating the effectiveness and rationality of the proposed GSC module in weakly supervised semantic segmentation tasks.

Table 5 illustrates the impact of different attention path designs in the RCSC module on model performance, serving to validate the effectiveness of the individual components of Spatial Attention (SA), Channel Attention (CA), and Residual Context Modulation (RCM). The experimental results show that each attention branch can enhance the model's semantic representation capabilities and segmentation performance to varying degrees. Compared to the baseline model's 74.3% mIoU, 84.0% Precision, and 85.7% Recall, introducing only SA improved model performance to 75.2% mIoU, with Precision and Recall reaching 85.3% and 87.7%, respectively. This demonstrates that spatial attention effectively enhances responses to target regions and improves spatial localization capabilities. Upon further integration of CA, the model's Precision improved to 85.5%, indicating that modeling channel-specific semantic dependencies helps enhance feature discrimination and contextual semantic representation. However, relying solely on spatial and channel attention remains insufficient for fully modeling long-range contextual relationships in complex scenes. Therefore, after further introducing RCM into the complete RCSC module, the model achieved optimal performance, with Precision, Recall, and mIoU reaching 85.6%, 87.8%, and 75.4%, respectively. Compared to the baseline model, the complete module achieved a 1.1% improvement in mIoU, indicating that residual context-semantic calibration effectively integrates spatial information with channel-semantic relationships, further enhancing the contextual consistency and boundary integrity of the target region while suppressing noise interference in complex backgrounds. The overall experimental results fully validate the effectiveness and complementarity of the various components within the proposed RCSC module.

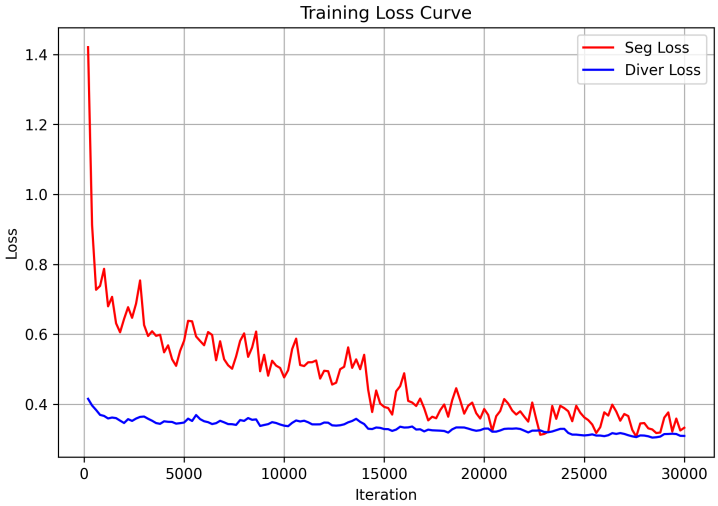


Fig. 5 Training loss curves of HCSC-Net. Both losses decrease steadily during training, indicating stable optimization and convergence.

4.4 Visualization results

As shown in Figure 5, throughout the training process, both the segmentation loss and the divergence loss exhibit a steady and continuous downward trend, demonstrating the effectiveness and optimization stability of the proposed HCSC-Net framework. In the early stages of training, the segmentation loss rapidly decreased from a relatively high initial value, indicating that the model can quickly learn discriminative semantic representations under weak supervision. As training progressed, the loss curve gradually smoothed out and continued to decline steadily, without any significant fluctuations or signs of optimization instability. This suggests that the proposed hierarchical semantic calibration strategy effectively promotes robust feature learning and stable parameter updates. Meanwhile, the divergence loss consistently maintains a low and smooth convergence trajectory, further validating the effectiveness of the proposed semantic calibration mechanism in enhancing semantic consistency and contextual compactness during the optimization process. In the later stages of training, both losses stabilized at relatively low values, indicating that the proposed framework achieved satisfactory convergence behavior without any signs of optimization collapse or premature saturation. Overall, the convergence characteristics shown in Figure 8 demonstrate that the proposed HCSC-Net possesses strong training stability, reliable optimization capabilities, and robust semantic representation learning capabilities throughout the entire training process.

As illustrated in Fig 6, the mIoU performance of both the semantic pseudo labels and segmentation predictions exhibits a generally increasing and stable trend throughout the training process, demonstrating the effectiveness and convergence stability of the proposed HCSC-Net framework. Specifically, the segmentation prediction mIoU shows a continuous improvement from approximately 65% in the early training stage

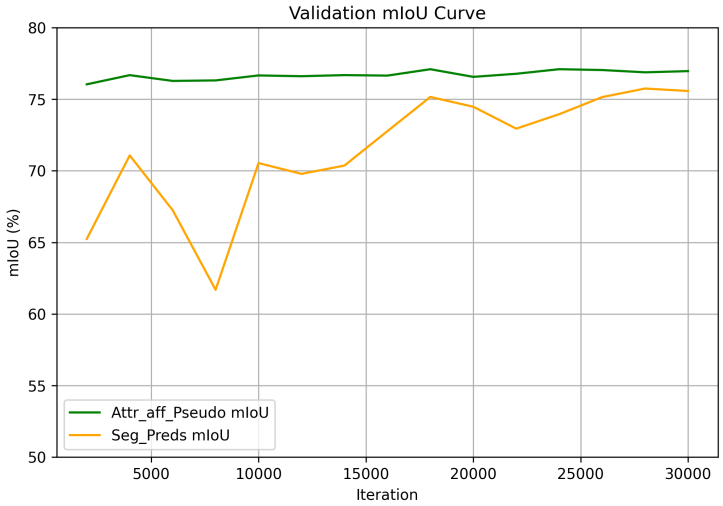


Fig. 6 Line chart of mIoU evolution. Both pseudo-label and segmentation mIoU values steadily increase during training, indicating stable convergence and improved segmentation performance.

to over 75% in the later optimization stage, indicating that the proposed framework progressively enhances dense semantic representation learning and foreground localization capability under weak supervision. Meanwhile, the mIoU of the attribute-affinity pseudo labels consistently maintains a relatively high level around 76%–77% during the entire training process, suggesting that the proposed hierarchical semantic calibration strategy provides stable and reliable semantic localization guidance for segmentation learning. In addition, both curves remain smooth without severe fluctuations or optimization collapse, which further verifies the robustness of the adopted optimization strategy and the strong semantic modeling capability of the proposed framework.

5 Conclusion

This paper proposes HCSC-Net, a novel single-stage, CLIP-based weakly supervised semantic segmentation framework designed to enhance dense semantic localization capabilities and semantic representation quality under weakly supervised conditions. To this end, we design the GSC module and the RCSC module. Through a hierarchical semantic calibration strategy, these modules enhance structural consistency before visual encoding and contextual semantic compactness after feature fusion, respectively. Specifically, the GSC module effectively suppresses noisy spatial responses and stabilizes local patch representations through adaptive group-wise structural calibration, while the RCSC module further enhances foreground region continuity and high-level semantic discrimination capabilities by leveraging residual contextual semantic interactions. Thanks to the synergistic effect of these two calibration mechanisms, HCSC-Net effectively alleviates issues such as activation region fragmentation, semantic ambiguity, and incomplete boundaries—problems prevalent in existing CLIP-based

WSSS methods—thereby generating more complete and discriminative dense semantic representations. Extensive experimental results demonstrate that the proposed method achieves performance superior to existing mainstream single-stage methods on both the PASCAL VOC 2012 and MS COCO 2014 datasets, without requiring additional training stages or complex post-processing steps. Furthermore, qualitative visualization results further validate that HCSC-Net can generate more continuous object activation regions and more compact semantic feature distributions in complex scenes. Nevertheless, our method still has certain limitations when handling extremely small objects and severely occluded scenes. In the future, we will further integrate adaptive multi-scale context modeling with stronger cross-modal semantic interaction mechanisms to continuously improve the model's robustness and generalization capabilities in complex environments.

Funding

This work was supported by the Key Scientific Research Project of Colleges and Universities in Henan Province (No. 25CY039, Xuezhuan Zhao); the Science and Technology Innovation Leading Talent Support Program of Henan Province, China (No. 254200510017, Lingling Li); the Research Programs of Henan Science and Technology Department (Nos. 252102210034 and 252102220053, Xiaoyan Shao and Peng Lu); the Henan Province Key Research and Development Project (No. 231111212000, Lingling Li); the Aviation Science Foundation of China (No. 20230001055002, Xuezhuan Zhao); the Henan Center for Outstanding Overseas Scientists (No. GZS2022011, Lingling Li); the National Natural Science Foundation of China, Youth Program (No. 62202434, Ying Duan); the Open Fund Project of Henan General Aviation Technology Key Laboratory (No. ZHKF-240205, Research on Multi-source Trustworthy Retrieval-augmented Generation Technology for Large Models in the Aviation Field, Changchang Zeng); and the 2025 Graduate Education Innovation Program of Zhengzhou University of Aeronautics (Nos. 2025CX121, 2025CX122, and 2025CX123).

Ethics declaration

Not applicable.

Conflict of interest

The authors declare that they have no conflict of interest.

Data availability

The datasets analyzed during the current study are publicly available from the corresponding public repositories.

References

[1] Chen, L., Lei, C., Li, R., Li, S., Zhang, Z., Zhang, L.: Fpr: False positive rectification for weakly supervised semantic segmentation. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 1108–1118 (2023)

[2] Zhang, B., Xiao, J., Jiao, J., Wei, Y., Zhao, Y.: Affinity attention graph neural network for weakly supervised semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **44**(11), 8082–8096 (2021)

[3] Wei, Y., Liang, X., Chen, Y., Jie, Z., Xiao, Y., Zhao, Y., Yan, S.: Learning to segment with image-level annotations. Pattern Recognit. (PR) **59**, 234–244 (2016)

[4] Xu, L., Bennamoun, M., Boussaid, F., Laga, H., Ouyang, W., Xu, D.: Mctformer+: Multi-class token transformer for weakly supervised semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) (2024)

[5] Zhang, B., Xiao, J., Jiao, J., Wei, Y., Zhao, Y.: Frozen clip: A strong backbone for weakly supervised semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3796–3806 (2024)

[6] Xie, J., Hou, X., Ye, K., Shen, L.: Clims: Cross language image matching for weakly supervised semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4483–4492 (2022)

[7] Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4981–4990 (2018)

[8] Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1713–1721 (2015)

[9] Yang, Z., Meng, Y., Fu, K., Wang, S., Song, Z.: Tackling ambiguity from perspective of uncertainty inference and affinity diversification for weakly supervised semantic segmentation. arXiv preprint arXiv:2404.08195 (2024)

[10] Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 1635–1643 (2015)

[11] Lee, J., Yi, J., Shin, C., Yoon, S.: Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2643–2652 (2021)

[12] Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proc. IEEE Conf. Comput. Vis.

Pattern Recognit. (CVPR), pp. 3159–3167 (2016)

[13] Vernaza, P., Chandraker, M.: Learning random-walk label propagation for weakly-supervised semantic segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 7158–7166 (2017)

[14] Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision. In: Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 549–565 (2016)

[15] Xu, Z., Tang, F., al., Z.C.: Toward modality gap: Vision prototype learning for weakly-supervised semantic segmentation with clip. In: Proc. AAAI Conf. Artif. Intell. (AAAI), vol. 39, pp. 9023–9031 (2025)

[16] Lin, Y., Chen, M., Zhang, K., al.: Tagclip: A local-to-global framework to enhance open-vocabulary multi-label classification of clip without training. In: Proc. AAAI Conf. Artif. Intell. (AAAI), vol. 38, pp. 3513–3521 (2024)

[17] Wu, Y., Ye, X., Yang, K., Li, J., Li, X.: Dupl: Dual student with trustworthy progressive learning for robust weakly supervised semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3534–3543 (2024)

[18] Yang, Z., Fu, K., Duan, M., Qu, L., Wang, S., Song, Z.: Separate and conquer: Decoupling co-occurrence via decomposition and representation for weakly supervised semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3606–3615 (2024)

[19] Rao, Y., Zhao, W., al., G.C.: Denseclip: Language-guided dense prediction with context-aware prompting. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 18082–18091 (2022)

[20] Ding, Z., Wang, J., Tu, Z.: Open-vocabulary universal image segmentation with maskclip. arXiv preprint arXiv:2208.08984 (2022)

[21] Lin, Y., Chen, M., Wang, W., Wu, B., Li, K., Lin, B., Liu, H., He, X.: Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. arXiv preprint arXiv:2212.09506 (2022)

[22] Du, Y., Fu, Z., Liu, Q., Wang, Y.: Weakly supervised semantic segmentation by pixel-to-prototype contrast. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4320–4329 (2022)

[23] Lee, J., Choi, J., Mok, J., Yoon, S.: Reducing information bottleneck for weakly supervised semantic segmentation. In: Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 27408–27421 (2021)

[24] Everingham, M., Eslami, S.M.A., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. Int. J. Comput. Vis. (IJCV) **111**(1), 98–136 (2015)

[25] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 740–755 (2014)

[26] Hariharan, B., Arbelaez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 991–998 (2011)

[27] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[28] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., al., S.G.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Int. Conf. Learn. Represent. (ICLR) (2021)

[29] Jang, S., Yun, J., al., J.K.: Dial: Dense image-text alignment for weakly supervised semantic segmentation. In: Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 248–266 (2024)

[30] Jiang, P.-T., Yang, Y., Hou, Q., Wei, Y.: L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 16886–16896 (2022)

[31] Zhou, T., Zhang, M., al., F.Z.: Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4299–4309 (2022)

[32] Cheng, Z., Qiao, P., Li, K., Li, S., Wei, P., Ji, X., Yuan, L., Liu, C., Chen, J.: Out-of-candidate rectification for weakly supervised semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 23673–23684 (2023)

[33] Rong, S., Tu, B., Wang, Z., Li, J.: Boundary-enhanced co-training for weakly supervised semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 19574–19584 (2023)

[34] Yoon, S.-H., Kwon, H., Kim, H., Yoon, K.-J.: Class tokens infusion for weakly supervised semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3595–3605 (2024)

[35] Zhao, X., Yang, Z., Dai, T., Zhang, B., Xiao, J.: Psdpm: Prototype-based secondary discriminative pixels mining for weakly supervised semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3437–3446 (2024)

[36] Tang, F., Xu, Z., Qu, Z., Feng, W., Jiang, X., Ge, Z.: Hunting attributes: Context prototype-aware learning for weakly supervised semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3324–3334 (2024)

[37] Ru, L., Zhan, Y., al., B.Y.: Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 16846–16855 (2022)

[38] Rossetti, S., Zappia, D., Sanzari, M., Schaerf, M., Pirri, F.: Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation. In: Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 446–463 (2022)

[39] Ru, L., Zheng, H., al., Y.Z.: Token contrast for weakly-supervised semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3093–3102 (2023)

[40] Xu, X., Zhang, P., al., W.H.: Weakly supervised semantic segmentation via progressive confidence region expansion. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 9829–9838 (2025)

[41] Chen, Z., Wang, T., Wu, X., Hua, X.-S., Zhang, H., Sun, Q.: Class re-activation maps for weakly-supervised semantic segmentation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 969–978 (2022)

[42] Chen, Z., Sun, Q.: Extracting class activation maps from non-discriminative features as well. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3135–3144 (2023)

[43] Zhao, X., Tang, F., Wang, X., Xiao, J.: Sfc: Shared feature calibration in weakly supervised semantic segmentation. In: Proc. AAAI Conf. Artif. Intell. (AAAI), pp. 7525–7533 (2024)

[44] Murugesan, B., Hussain, R., Bhattacharya, R., Ayed, I.B., Dolz, J.: Prompting classes: Exploring the power of prompt class learning in weakly supervised semantic segmentation. In: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), pp. 291–302 (2024)