2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Multi-state Consistency Visual Language Model Combine Wavelet Transform for Weakly Supervised Robot Image Segmentation

Feng Xiao[1], Peihua Han[1], Guoyuan Li[1] and Houxiang Zhang[1]

***Abstract*— Robotic visual segmentation is essential for enabling robots to operate in complex environments. Although supervised methods have achieved remarkable progress, their dependence on dense annotations hinders scalability. Weakly supervised semantic segmentation (WSSS) alleviates this issue but suffers from sparse supervision, leading to noisy pseudo-labels and boundary errors. Large visual models (LVMs), pre-trained on diverse data, provide rich semantic priors that can strengthen weak supervision and address these limitations. To this end, we designed a dual-branch architecture, introducing two large pre-trained models with complementary characteristics. We align the feature spaces of the two branches through consistency learning to alleviate the representation differences and weakly supervised noise problems caused by cross-domain migration, thereby obtaining more robust and fine-grained semantic features. Furthermore, to effectively restore spatial details and improve the quality of segmentation boundaries, we introduce a wavelet transform in the decoder. Wavelet decomposition can simultaneously capture low-frequency global information and high-frequency local details at multiple scales, allowing the model to enhance spatial restoration capabilities while maintaining semantic consistency. Experimental results show that our method improves the performance by 7.7% compared with the state-of-the-art methods in WSSS.**

I. Introduction

Robotic visual segmentation has become one of the most critical foundations for enabling intelligent robots to interact with complex and dynamic real-world environments. From autonomous navigation and object manipulation to human–robot interaction [1], segmentation serves as the prerequisite for decision-making and control. In recent years, deep learning has significantly advanced visual segmentation in robotics, driven by large-scale annotated datasets and convolutional and transformer-based architectures [2], [3], [4], [5], [6]. However, despite this progress, conventional supervised learning paradigms face severe limitations when extended to robotic applications, where data annotation is often prohibitively costly, task-specific, and environmentally constrained. Weakly supervised visual segmentation (WSSS) [7], [8], which relies on limited or noisy labels instead of fully annotated datasets, has emerged as a promising solution to reduce dependency on exhaustive human labeling. Nevertheless, weak supervision in robotics still struggles to achieve robust performance under diverse and unstructured real-world conditions.

[1]Feng Xiao, Peihua Han, Guoyuan Li, and Houxiang Zhang are with the Department of Ocean Operations and Civil Engineering, Faculty of Engineering, Norwegian University of Science and Technology. {feng.xiao, peihua.han, guoyuan.li, hozh}@ntnu.no

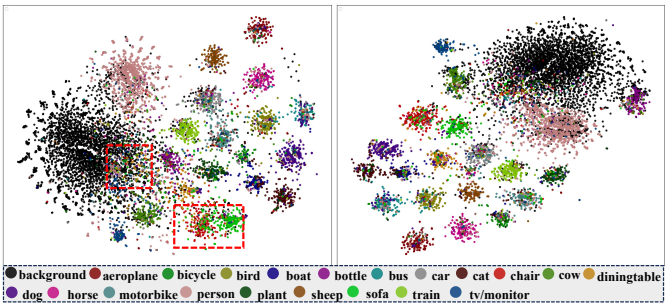

Fig. 1. The left figure shows the feature distribution of the baseline method, and the right figure shows the feature distribution of our method. Different colors represent different categories, and we can see that our method achieves more compact and discriminative clustering between categories. The red dashed line indicates an incorrect clustering.

With the rapid development of large visual models (LVMs) pretrained on massive, diverse, and multimodal data, opportunities have arisen to transfer their generalization ability and semantic richness to robotic segmentation tasks. LVMs such as Segment Anything Model (SAM)[9], and DINO [10], [11] have shown unprecedented capability in capturing both global semantics and fine-grained structural cues. These characteristics are particularly valuable for robots, which often face tasks in dynamic, cluttered, or partially observable environments. Yet, a direct adoption of LVMs in robotic visual segmentation remains challenging. The pretrained objectives of LVMs may not align with robot-specific requirements; computational overheads can be prohibitive for onboard deployment.

Weakly supervised approaches provide a promising alternative but continue to face several unresolved challenges. First, although pretrained on massive internet-scale datasets, models often suffer from domain adaptation difficulties when applied to robotic data characterized by unusual viewpoints, occlusions, and sensor-specific noise. This severe domain shift exacerbates the noise in weakly supervised settings, making robust feature adaptation highly difficult. Second, robotic tasks demand both holistic scene understanding and fine-grained recognition (e.g., grasp points or obstacle edges), whereas most pretrained models primarily emphasize either global semantics or local textures, leading to insufficient semantic granularity. As shown in Fig. 1, the baseline method based on single branch CLIP on the left suffers from obvious misclassification, while our method can significantly improve the performance. Furthermore, standard decoding processes struggle to accurately recover high-frequency spa-

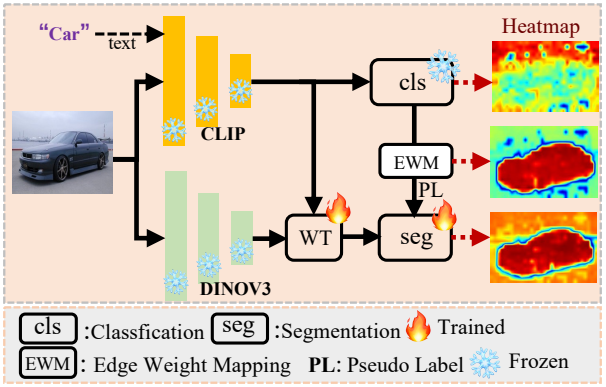

Fig. 2. Our method achieves SOTA performance by only using edge weight mapping and wavelet transform in a single-stage training process.

tial details from these models, often resulting in blurred segmentation boundaries.

To address these issues, we propose a dual-branch large visual model framework designed for weakly supervised robotic segmentation. As shown in Fig. 2, specifically, our method integrates two complementary pretrained LVMs: one branch focuses on capturing global semantic context, while the other emphasizes local structural details. To alleviate representation gaps and mitigate noise introduced by weak supervision, we employ consistency learning to align the feature spaces of the two branches, thereby obtaining more robust and fine-grained semantic features. Furthermore, to enhance spatial restoration and segmentation quality, we introduce a wavelet-based decoder, where multiscale wavelet decomposition jointly captures low-frequency global information and high-frequency local details, leading to sharper and more semantically consistent predictions.

Our main contributions are summarized as follows:

1) A dual-branch architecture is proposed to integrate two complementary pretrained LVMs, enabling concurrent segmentation of global semantic context and local structural cues.
2) A weakly supervised adaptation strategy based on consistency learning is developed to suppress cross-domain migration noise and improve robustness under limited or noisy annotations.
3) A wavelet-transform-based decoder is introduced to better recover spatial details, yielding sharper segmentation boundaries while preserving semantic consistency.
4) Comprehensive experiments show that our method achieves high performance in weakly supervised settings, while also maintaining computational efficiency for real-time robotic applications.

II. Related Work

A. Traditional Supervised Learning

Conventional semantic segmentation methods under supervised learning usually depend on large-scale datasets with pixel-level annotations, such as PASCAL VOC, MS COCO, and Cityscapes. Earlier studies mainly employed fully convolutional networks (FCNs) [12], which reformulated image classification architectures into dense prediction frameworks by substituting fully connected layers with convolutional ones. Building on FCNs, various network architectures have been proposed to enhance feature extraction and multi-scale representation. For example, U-Net [13] and DeepLab series [14] introduce skip connections, atrous convolution, and pyramid pooling, respectively, to improve the ability to capture fine-grained details and global context. Although these methods can reach high accuracy, their practical application is hindered by the heavy reliance on pixel-level annotations, which are costly, labor-intensive, and difficult to obtain in complex domains such as medical diagnosis or underwater vision. Furthermore, the effectiveness of these approaches strongly depends on the diversity and quality of training datasets, restricting their adaptability to new situations with limited labeled samples. These limitations have driven increasing attention toward weakly supervised and unsupervised strategies.

B. Large Vision Model Weakly Supervised Semantic Segmentation

Recently, LVMs pretrained on massive datasets with self-supervised or weakly supervised objectives have demonstrated strong generalization and semantic understanding across diverse visual tasks. Representative models such as CLIP [9], DINOv3 [15], [11], and SAM [16] provide rich feature representations that can be adapted to downstream segmentation with minimal supervision. In WSSS, LVMs are widely employed to lower the reliance on dense annotations by utilizing weaker supervision signals, including image-level labels, bounding boxes, or class activation maps (CAMs). For example, pretrained Transformer architectures are capable of producing informative attention maps, which can be exploited as pseudo-labels to guide the training of segmentation models. Similarly, SAM's promptable segmentation capability enables flexible region extraction without requiring dense supervision. Recent studies explore dual-branch or hybrid frameworks that combine different LVMs to balance global semantic understanding and local detail capture. Such approaches demonstrate the potential of LVM-driven WSSS in addressing complex visual environments where traditional supervised methods struggle. However, challenges remain in terms of precise boundary delineation, handling class imbalance, and adapting pretrained representations to domain-specific tasks, which motivates further research in integrating LVMs priors with weakly supervised segmentation frameworks.

III. Methodology

A. Overview Framework

As illustrated in Fig. 3, the framework attains WSSS by exploiting the complementary strengths of vision-language models and self-supervised encoders. The pipeline consists of four main components: (1) a CLIP text encoder for semantic prompt embedding, (2) dual visual backbones based

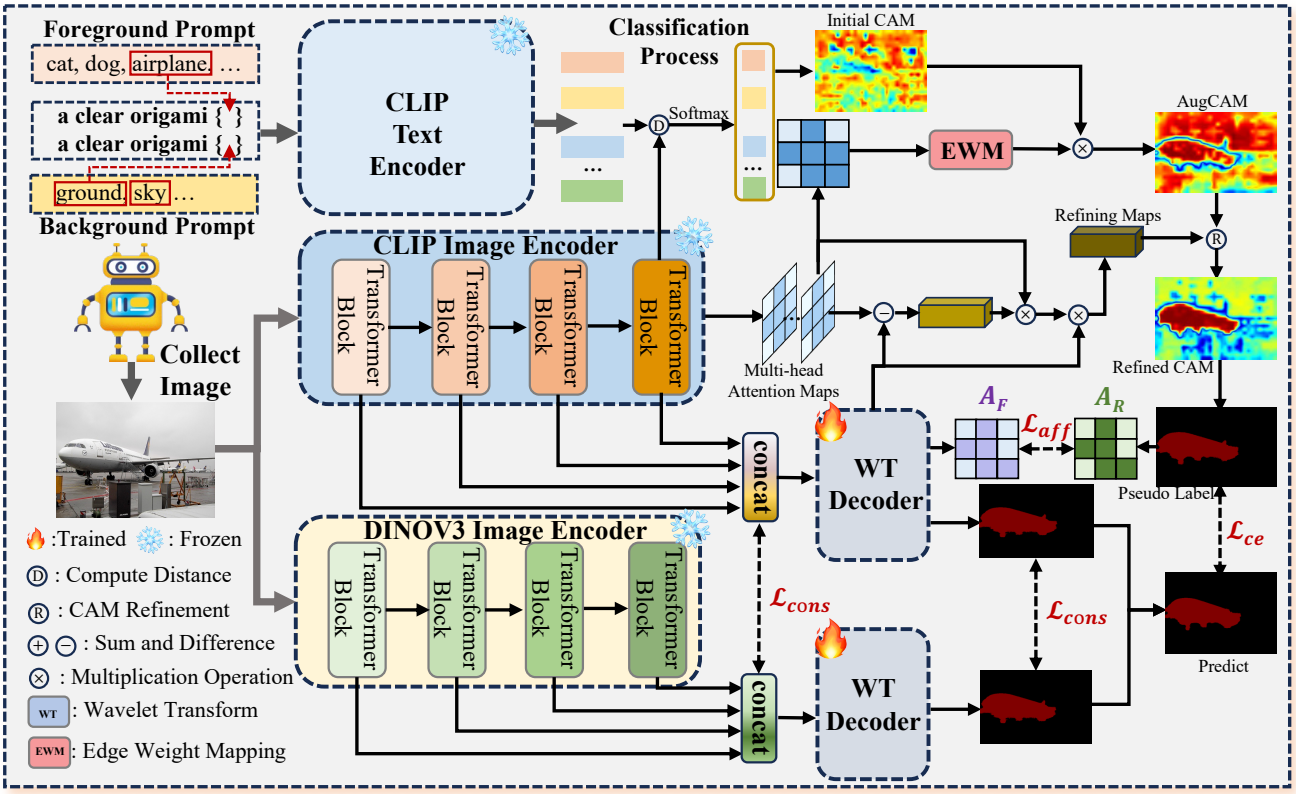

Fig. 3. This diagram outlines our framework, which consists of four key components: two pre-trained CLIP and DINO backbones, a classification process for initial CAM generation, a Wavelet Transform decoder for segmentation predictions, and a branch for refining the initial CAM to generate training pseudo-labels.

on CLIP and DINOv3, (3) a Wavelet Transform (WT) decoder for segmentation predictions, and (4) a refinement branch for generating high-quality pseudo labels. First, textual prompts describing potential foreground and background categories are encoded by the frozen CLIP text encoder. Simultaneously, the input image is fed into two parallel image encoders, namely the CLIP image encoder and the DINOv3 encoder. Both encoders extract hierarchical visual features via multiple transformer blocks. A classification process computes the similarity between the text and image embeddings, producing class activation maps (CAMs). The initial CAM serves as the foundation for the localization of semantic regions. To enhance localization accuracy, an Edge Weight Mapping (EWM) module and a CAM refinement process are employed to generate an augmented and refined CAM, which acts as a pseudo-label supervision signal. In parallel, the multi-level features extracted by CLIP and DINOv3 encoders are passed through a concat block before being decoded by the WT decoder. The wavelet transform allows the decoder to capture both global context and fine-grained details, while the inverse wavelet transform ensures effective reconstruction of semantic regions. The refined CAMs are then integrated with the WT decoder outputs, enabling the network to generate segmentation predictions supervised by cross-entropy loss. Through this collaborative design, the framework leverages both language-guided semantic priors and visual self-supervised representations, resulting in robust and high-quality segmentation under weak supervision.

B. Dual-branch Visual Large Encoder

To extract complementary semantic and structural features, we design a dual-branch encoder consisting of CLIP and DINOv3. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the CLIP text encoder $\mathcal{E}_{txt}$ encodes foreground and background prompts $T = \{t_f, t_b\}$ into text embeddings:

$$z_{txt} = \mathcal{E}_{txt}(T). \tag{1}$$

Meanwhile, the CLIP image encoder $\mathcal{E}_{img}^{CLIP}$ extracts semantic embeddings:

$$z_{clip} = \mathcal{E}_{img}^{CLIP}(I). \tag{2}$$

Their similarity is computed by

$$s = \text{softmax}(z_{txt}^{\top} z_{clip}), \tag{3}$$

and the initial CAM is obtained as

$$M_{clip} = \sum_k s_k \cdot F_k, \tag{4}$$

where F_k denotes the k-th feature map. To further enhance the quality of localization maps, we introduce an EWM. The EWM computes edge-aware weights W_{edge} based on

structural cues of the input, which are then multiplied with the initial CAM to emphasize boundary-consistent regions:

$$M_{aug} = M_{clip} \odot W_{edge}. \tag{5}$$

Here $\odot$ denotes element-wise multiplication. This step refines the initial CAM by suppressing ambiguous activations and highlighting more precise object boundaries. At the same time, the DINOv3 encoder extracts structural features:

$$z_{dino} = \mathcal{E}_{img}^{DINO}(I). \tag{6}$$

Providing rich multi-view representations that are subsequently decoded to generate refined masks and pseudo labels.

C. Wavelet Transform Decoder

Conventional decoders based on convolution or naive upsampling often fail to preserve boundary sharpness and fine-grained consistency, especially when dealing with complex object contours. To address this issue, we design a Wavelet Transform (WT)-based decoder that operates on both CLIP and DINOv3 branch features, enabling multi-scale decomposition and reconstruction in both spatial and frequency domains.

Specifically, we apply a discrete wavelet transform (DWT) separately to z_{clip} and z_{dino}, decomposing each feature map into one low-frequency approximation component A and three high-frequency detail components H, V, D:

$$\{A^{clip}, H^{clip}, V^{clip}, D^{clip}\} = \mathrm{DWT}(z_{clip}), \tag{7}$$

$$\{A^{dino}, H^{dino}, V^{dino}, D^{dino}\} = \mathrm{DWT}(z_{dino}), \tag{8}$$

where A captures global semantic structures, while H, V, D encode horizontal, vertical, and diagonal edge and texture information.

To enhance discriminative capacity, each sub-band component is refined through a learnable transformation function $\mathcal{F}$:

$$A' = \mathcal{F}(A), \quad H' = \mathcal{F}(H), \quad V' = \mathcal{F}(V), \quad D' = \mathcal{F}(D). \tag{9}$$

The refined components are then reconstructed using the inverse wavelet transform (IWT) to obtain high-resolution feature maps for each branch:

$$F_{wt}^{clip} = \mathrm{IWT}(A'^{clip}, H'^{clip}, V'^{clip}, D'^{clip}), \tag{10}$$

$$F_{wt}^{dino} = \mathrm{IWT}(A'^{dino}, H'^{dino}, V'^{dino}, D'^{dino}). \tag{11}$$

This adaptive weighting allows the decoder to dynamically emphasize global context or local details depending on the characteristics of the input image. Finally, the decoder outputs pixel-wise prediction probabilities:

$$\hat{Y} = (F_{wt}^{dino} + F_{wt}^{clip})/2. \tag{12}$$

The decoded feature maps F_{wt}^{clip} and F_{wt}^{dino} are further utilized in two ways: (1) generating pseudo labels supervised by the cross-entropy loss L_{ce}, and (2) enforcing cross-branch consistency via the consistency loss L_{cons}, which aligns the semantic and structural outputs of the two branches.

Compared with conventional decoders, the proposed WT decoder offers three major advantages: 1. **Semantic-structural separation**: DWT explicitly decouples low-frequency semantics and high-frequency edge details, facilitating fine-grained structural learning. 2. **Cross-branch alignment**: The consistency loss L_{cons} ensures that CLIP and DINO features remain aligned at both global and local levels. 3. **Detail preservation and boundary enhancement**: Multi-scale wavelet reconstruction mitigates over-smoothing during upsampling, producing sharper boundaries and more accurate object contours.

D. Loss Function

An essential component of our framework is the loss function, which guides the optimization process and ensures that the predictions are both semantically consistent and structurally accurate. Since our framework integrates multi-branch encoders and wavelet-based decoding, the design of the loss function must account for multiple aspects of learning: accurate classification and consistency with refined pseudo-labels. To achieve this, we construct a hybrid loss that combines cross-entropy loss and consistency regularization.

a) Cross-Entropy Loss: The primary objective in semantic segmentation is to classify each pixel into one of C predefined categories. We adopt the standard pixel-wise cross-entropy (CE) loss as the baseline supervisory signal:

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log \hat{y}_{i,c}, \tag{13}$$

here N denotes the overall pixel count. $y_{i,c} \in \{0,1\}$ indicates the ground-truth assignment of pixel i to class c, and $\hat{y}_{i,c}$ stands for the network's output probability for that specific class. The CE loss penalizes misclassification at the pixel level, ensuring global semantic correctness.

b) Affinity Consistency Loss: In addition to pixel-wise supervision, we further exploit feature affinity to capture structural relationships between pixels. Specifically, we construct an affinity map A_F from the decoder feature map, where each entry encodes the similarity between a pair of pixels. To guide this affinity map, we convert the pseudo labels M_p into the corresponding label affinity map A_R, where $A_R(i,j) = 1$ if pixels i and j belong to the same class, and 0 otherwise. The consistency between A_F and A_R is enforced using a cross-entropy loss:

$$\mathcal{L}_{aff} = \mathcal{L}_{ce}\left(A_F, A_R\right). \tag{14}$$

This affinity-based constraint enables the model to learn more reliable pairwise relationships, which in turn improve the quality of pseudo labels and enhance structural consistency across object regions.

c) Multi-state Consistency Loss: Beyond supervised learning, we encourage mutual alignment between the two branches. Specifically, the CLIP branch prediction $\hat{Y}^{clip}$ and the DINO branch prediction $\hat{Y}^{dino}$ serve as reciprocal supervision signals. We employ an L_1 consistency loss to

enforce agreement between them:

$$\mathcal{L}_{cons} = \frac{1}{N} \sum i = 1^{N} |\hat{y}_i^{clip} - \hat{y}_i^{dino}|_1, \tag{15}$$

where $\hat{y}_i^{clip}$ and $\hat{y}_i^{dino}$ denote the pixel-level predictions from the two branches. This consistency constraint reduces representation discrepancy, mitigates branch-specific noise, and enhances the robustness of the learned features.

d) Total Objective: Finally, the total loss function is defined as a weighted sum of the three components:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{aff} + \lambda_2 \mathcal{L}_{cons}, \tag{16}$$

where λ_1 and λ_2 denote hyperparameters used to control the trade-off between the supervised loss and the consistency regularization term.

IV. Experiment

A. Datasets

We evaluate our framework on two widely used benchmarks: PASCAL VOC2012 and MS COCO 2014. **PASCAL VOC2012:** This dataset consists of 20 object classes plus background. We use the augmented version with additional annotations from [17], resulting in 10,582 training, 1,449 validation, and 1,456 test images. Its high intra-class variability and complex object boundaries make it a challenging benchmark for assessing both semantic accuracy and boundary precision. **MS COCO 2014:** COCO is much larger, with 80 categories, 82,783 training images, and 40,504 validation images. Each image contains multiple annotated objects, leading to increased scene complexity, severe occlusions, and significant scale variation. Evaluating on COCO tests the robustness and generalization of our method in more diverse and realistic settings.

Benchmarking various WSSS frameworks using the COCO 2014 dataset.

B. Implementation Details

Our framework is implemented in PyTorch and trained on two NVIDIA A6000 GPUs. We use the AdamW optimizer with an initial learning rate of 2×10^{-5}, weight decay of 1×10^{-2}, and a cosine annealing schedule for gradual learning rate adjustment. We follow previous work [18] and use mIoU as the evaluation metric. We adopt a batch size of 8, and the training process lasts for 80 epochs on VOC2012 and 120 epochs on COCO 2014. The dual-branch encoder is initialized with pretrained CLIP and DINOv3 models, while the wavelet-based decoder and refinement modules are learned from scratch. The pre-trained weights for the two encoders are VIT-B-16. For evaluation, we follow the common practice: VOC2012 results are obtained via the official server, and COCO 2014 results are reported on the validation set according to the standard protocol. This configuration guarantees a fair comparison with existing state-of-the-art methods. The hyperparameters λ_1 and λ_2 of our loss function are set to 0.1 and 0.2 to achieve the best performance.

TABLE I
QUANTITATIVE RESULTS OF DIFFERENT WSSS APPROACHES ON THE PASCAL VOC 2012 DATASET.

Method	Backbone	Supervised	val	test
Multi-stage WSSS				
Mat-label$_{\text{ICCV'23}}$	ResNet101	I+S	73.3	**74.0**
ESOL$_{\text{NeurIPS'22}}$	ResNet101	I	69.9	69.3
VML$_{\text{IJCV'23}}$	ResNet101	I	70.6	70.7
AETF$_{\text{ECCV'22}}$	ResNet38	I	70.9	71.7
MCTformer$_{\text{CVPR'22}}$	ViT+Res38	I	70.4	70.0
CDL$_{\text{IJCV'23}}$	ResNet101	I	72.4	72.2
ACR$_{\text{CVPR'23}}$	ViT	I	72.4	72.4
BECO$_{\text{CVPR'23}}$	MIT-B2	I	73.7	73.5
CLIP-ES$_{\text{CVPR'23}}$	ViT+Res101	I+L	73.8	73.9
FPR$_{\text{ICCV'23}}$	ResNet101	I	70.0	70.6
USAGE$_{\text{ICCV'23}}$	ResNet38	I	71.9	72.8
Single-stage WSSS				
SLRNet$_{\text{IJCV'22}}$	ResNet38	I	67.2	67.6
DuPL$_{\text{CVPR'24}}$	ViT-B	I	73.3	72.8
TSCD$_{\text{AAAI'23}}$	MIT-B1	I	67.3	67.5
ToCo$_{\text{CVPR'23}}$	ViT	I	71.1	72.2
MoRe$_{\text{AAAI'25}}$	ViT	I	76.4	75.0
SeCO$_{\text{CVPR'24}}$	ViT	I	74.0	73.8
WeCLIP$_{\text{CVPR'24}}$	ViT	I+L	74.9	75.2
ExCEL$_{\text{CVPR'25}}$	ViT-B	I+L	77.2	77.3
VPL$_{\text{AAAI'25}}$	ViT-B	I+L	78.0	<u>77.8</u>
Ours	ViT-B	I+L	**82.0**	**82.6**

C. Qualitative results analysis

In this paper, we selected some current SOTA performance methods, including Mat-label [19], MCTFormer [20], SIPE [21], ESOL [22], FRP [23], CDL [24], ACR [25], BECO [26], USAGE [27], CLIP-ES [28], SLRNet [29], DuPL [30], TSCD [31], ToCo [32], WECLIP [17], MoRe [33], SeCO [34], ExCEL [35], VPL [36], and DIAL [37].

Results on PASCAL VOC2012: Table I, Fig.4, and Fig.5 present the comparison between our framework and recent WSSS methods on the PASCAL VOC 2012 benchmark. Bold text represents the best performance, and underlined text represents the second best. For clarity, existing approaches are grouped into multi-stage and single-stage pipelines. Multi-stage strategies, often relying on extra refinement modules or saliency/seed generation, achieve competitive results (e.g., CLIP-ES with 73.8% mIoU) but generally involve complicated training and limited scalability. In contrast, our approach follows the single-stage paradigm, offering a more concise design without external post-processing. It sets a new record for single-stage WSSS: 82.6% mIoU on the test set, surpassing ExCEL (77.3%) and VPL (77.8%) by nearly +5%. On the validation set, our model also achieves 82.0%, outperforming previous methods such as WeCLIP (74.9%) and MoRe (76.4%). These results confirm that the proposed dual-branch architecture effectively balances global semantics and local details, delivering SOTA accuracy while

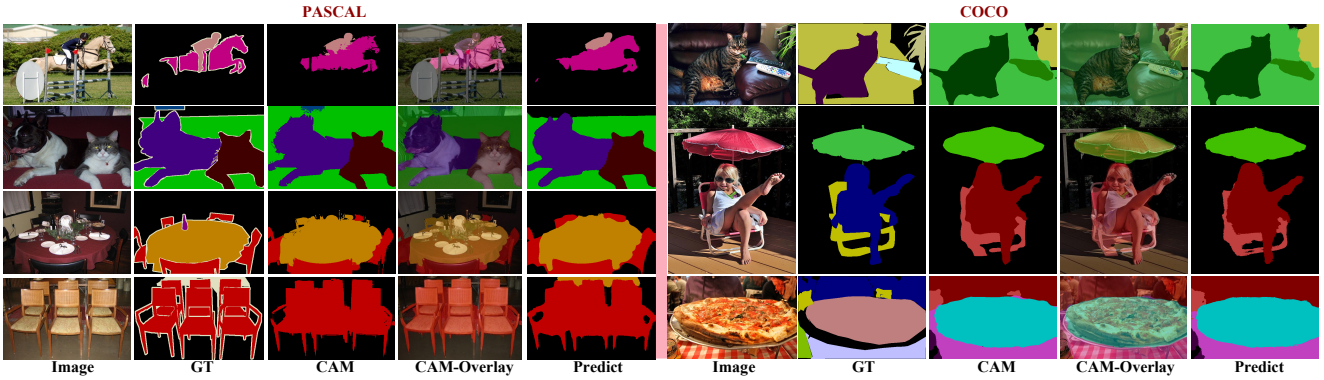

Fig. 4. Visualization of segmentation results. Columns from left to right represent the input image, ground truth (GT), class activation map (CAM), CAM overlaid on the image, and the final predicted mask. These comparisons highlight the ability of our approach to refine rough CAM outputs and generate predictions more consistent with GT.

TABLE II
BENCHMARKING VARIOUS WSSS FRAMEWORKS USING THE COCO 2014 DATASET.

Method	Backbone	Supervised	mIoU (%)
	Multi-stage WSSS		
$\text{MCTformer}_{\text{CVPR'22}}$	ViT+Res38	I	42.0
$\text{ESOL}_{\text{NeurIPS'22}}$	ResNet101	I	42.6
$\text{SIPE}_{\text{CVPR'22}}$	ResNet38	I	43.6
$\text{FPR}_{\text{ICCV'23}}$	ResNet101	I	43.9
$\text{USAGE}_{\text{ICCV'23}}$	ResNet101	I	44.3
$\text{ACR}_{\text{CVPR'23}}$	ResNet38	I	45.3
$\text{CDL}_{\text{IJCV'23}}$	ResNet101	I	**45.5**
$\text{CLIP-ES}_{\text{CVPR'23}}$	ViT+Res101	I+L	45.4
$\text{BECO}_{\text{CVPR'23}}$	ViT	I	45.1
	Single-stage WSSS		
$\text{DuPL}_{\text{CVPR'24}}$	ViT-B	I	44.6
$\text{TSCD}_{\text{AAAI'23}}$	MIT-B1	I	40.1
$\text{ToCo}_{\text{CVPR'23}}$	ViT	I	42.3
$\text{WeCLIP}_{\text{CVPR'24}}$	ViT	I+L	47.1
$\text{DIAL}_{\text{ECCV'24}}$	ViT-B	I+L	44.4
$\text{MoRe}_{\text{AAAI'25}}$	ViT-B	I	47.4
$\text{ExCEL}_{\text{CVPR'25}}$	ViT-B	I+L	<u>49.3</u>
Ours	ViT-B	I+L	**56.0**

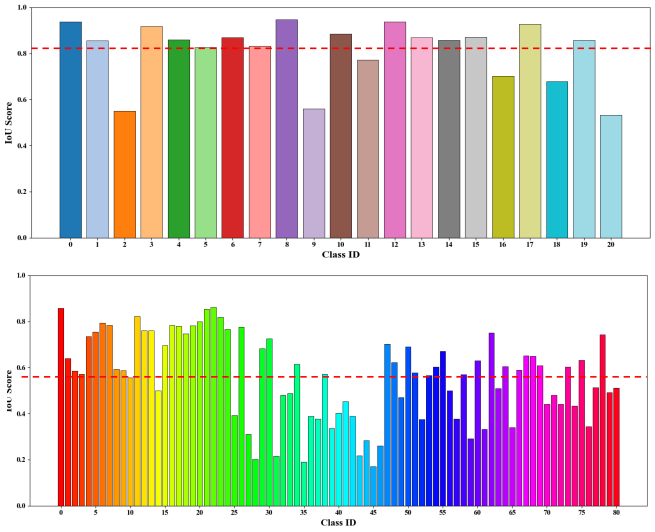

Fig. 5. Per-class IoU scores on two datasets. The red dashed line indicates the mIoU across all classes.

retaining simplicity and scalability.

Results on COCO2014: To assess the generalization ability of our approach, additional experiments are performed on the COCO 2014 benchmark, with the outcomes reported in Table II, Fig.4, and Fig.5. Using the same evaluation protocol as on VOC, we compare our method with both single-stage and multi-stage WSSS approaches. Within the multi-stage group, CDL and CLIP-ES achieve 45.5% and 45.4% mIoU, respectively, indicating competitive results yet with reliance on refinement. Despite their effectiveness, these approaches rely heavily on additional refinement steps or complex supervision strategies (e.g., saliency maps, class activation maps), which make them less efficient and harder to extend to large-scale scenarios. For single-stage approaches that are more concise and efficient, our proposed method establishes a new SOTA system. As shown in Table II, our method reaches 56.0% mIoU, significantly surpassing the previous best method ExCEL, which achieved 49.3%. This improvement of more than +7.7% demonstrates the strong capability of our method in handling complex and large-scale datasets like COCO 2014. Moreover, compared with other recent methods such as WeCLIP (47.1%) and MoRe (47.4%), our method consistently shows superior performance.

D. Quantitative results analysis

Fig. 4 shows qualitative examples produced by our WSSS framework. From left to right, each row displays the input image and the corresponding ground-truth (GT) annotation, followed by the segmentation predictions generated by our model and comparison methods. The third column shows the raw CAM, which provides coarse object localization but often lacks precise boundaries. To better visualize the localization quality, the fourth column overlays the CAM

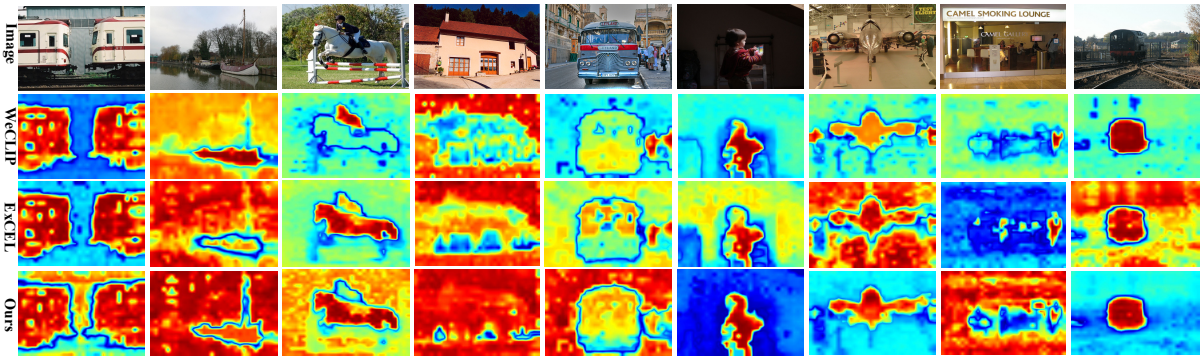

Fig. 6. Visual comparison across methods on representative samples. The first row presents the input image, while the subsequent rows (second to fourth) display the response heatmaps produced by different approaches.

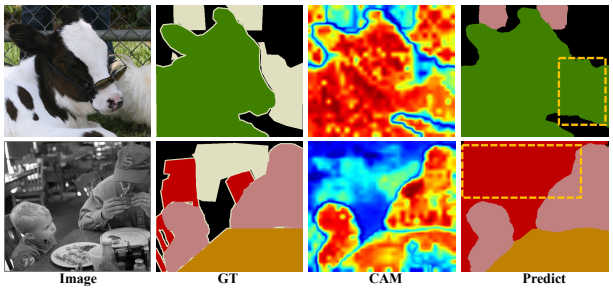

Fig. 7. Failure cases of our method. The yellow dotted box represents the mis-segmented area.

on the original image, revealing that the highlighted regions capture the general object areas but still contain noise and incomplete coverage. Finally, the last column displays the predicted segmentation masks generated by our method. Compared with the initial CAM, the predictions exhibit substantially improved object boundary delineation and region completeness, aligning more closely with the ground truth annotations. These results demonstrate the effectiveness of the proposed approach in refining CAM-based cues into high-quality segmentation outcomes. In addition, we demonstrate the performance of our method in Fig. 6, where our method produces more accurate localization and reduced background noise compared with other approaches.

TABLE III

PERFORMANCE COMPARISON OF DIFFERENT CONFIGURATIONS.

BASE-CLIP	DINO	WT	LOSS	mIoU
✓				43.0
✓		✓	✓	48.0
✓		✓		45.6
✓	✓	✓		55.2
✓	✓		✓	54.4
✓	✓	✓	✓	**56.0**

E. Ablation Study

Table III presents the ablation study evaluating the effectiveness of different components in our framework, including the BASE-CLIP encoder, DINO encoder, wavelet transform (WT), and the designed loss function. Starting from the baseline configuration with only BASE-CLIP, the model achieves an mIoU of 43.0. Incorporating WT improves performance to 48.0, while adding DINO further boosts mIoU to 54.4. When WT is used without DINO, the performance is 45.6, highlighting the necessity of complementary encoders. Finally, the full system, composed of BASE-CLIP, DINO, and WT, along with the proposed loss function, delivers the top result with an mIoU of 56.0, illustrating the robustness and synergy achieved through their integration.

F. Analysis of Failure Cases

Despite the overall effectiveness of our framework, several failure cases remain. As illustrated in Fig. 7, the first example demonstrates that our model struggles with fine-grained object boundaries when foreground and background textures are highly similar. In such cases, the segmentation output tends to blur object contours, leading to partial misclassification. The second example highlights difficulties in complex multi-object scenes with severe occlusions. Here, adjacent instances, such as human faces and surrounding objects, are confused, resulting in boundary leakage and fragmented predictions. These observations indicate that our method is sensitive to boundary ambiguity and occlusion, which are common yet challenging issues in real-world scenarios. Future work could address these limitations by incorporating boundary-aware refinement modules, leveraging multi-scale context more effectively, or introducing stronger instance-level priors.

V. CONCLUSION

We propose a dual-branch visual model for WSSS, integrating two pretrained LVMs to capture global semantics and local details. A consistency-based adaptation strategy reduces noise from WSSS, while a wavelet transform decoder improves spatial restoration and boundary precision. Experimental results show our approach outperforms existing methods in weakly supervised settings, with real-time deployment efficiency. Future challenges include domain shift, multimodal scalability, and long-term adaptability, which we aim to address using transfer learning, multimodal integration, and knowledge distillation.

REFERENCES

[1] A. Nekrasov, R. Zhou, M. Ackermann, A. Hermans, B. Leibe, and M. Rottmann, "Oodis: Anomaly instance segmentation and detection benchmark," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 2764–2771.

[2] H. Liu, C. Jia, F. Shi, X. Cheng, and S. Chen, "Scsegamba: lightweight structure-aware vision mamba for crack segmentation in structures," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 29 406–29 416.

[3] W. Tian, X. Cheng, G. Li, F. Shi, S. Chen, and H. Zhang, "A multilevel convolutional recurrent neural network for blade icing detection of wind turbine," *IEEE Sensors Journal*, vol. 21, no. 18, pp. 20 311–20 323, 2021.

[4] F. Xiao, R. Liu, X. Cheng, H. Zhang, J. Zhang, and Y. Jin, "Dual-branch semantic enhancement network joint with iterative self-matching training strategy for semi-supervised semantic segmentation," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2025.

[5] P. Han, G. Li, S. Skjong, and H. Zhang, "Directional wave spectrum estimation with ship motion responses using adversarial networks," *Marine Structures*, vol. 83, p. 103159, 2022.

[6] A. L. Ellefsen, S. Ushakov, V. Æsøy, and H. Zhang, "Validation of data-driven labeling approaches using a novel deep network structure for remaining useful life predictions," *IEEE Access*, vol. 7, pp. 71 563–71 575, 2019.

[7] B. Zhang, S. Yu, J. Xiao, Y. Wei, and Y. Zhao, "Frozen clip-dino: A strong backbone for weakly supervised semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[8] F. Xiao, J. Zhang, P. Han, S. Chen, and H. Zhang, "Wtclip: A wavelet-aware clip framework for boundary-refined weakly supervised semantic segmentation," *IEEE Transactions on Industrial Informatics*, 2026.

[9] A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[10] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 9630–9640.

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

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[14] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[15] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski, "DINOv3," 2025. [Online]. Available: https://arxiv.org/abs/2508.10104

[16] A. Kirillov, E. Mintun, N. Ravi *et al.*, "Segment anything," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023.

[17] B. Zhang, S. Yu, Y. Wei, Y. Zhao, and J. Xiao, "Frozen clip: A strong backbone for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3796–3806.

[18] X. Zhao, F. Tang, X. Wang, and J. Xiao, "Sfc: Shared feature calibration in weakly supervised semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7525–7533.

[19] C. Wang, R. Xu, S. Xu, W. Meng, and X. Zhang, "Treating pseudo-labels generation as image matting for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 755–765.

[20] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, and D. Xu, "Multi-class token transformer for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4310–4319.

[21] T. Wu, G. Gao, J. Huang, X. Wei, X. Wei, and C. H. Liu, "Adaptive spatial-bce loss for weakly supervised semantic segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 199–216.

[22] J. Li, Z. Jie, X. Wang, X. Wei, and L. Ma, "Expansion and shrinkage of localization for weakly-supervised semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 037–16 051, 2022.

[23] L. Chen, C. Lei, R. Li, S. Li, Z. Zhang, and L. Zhang, "Fpr: False positive rectification for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1108–1118.

[24] B. Zhang, J. Xiao, Y. Wei, and Y. Zhao, "Credible dual-expert learning for weakly supervised semantic segmentation," *International Journal of Computer Vision*, vol. 131, no. 8, pp. 1892–1908, 2023.

[25] H. Kweon, S.-H. Yoon, and K.-J. Yoon, "Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 11 329–11 339.

[26] S. Rong, B. Tu, Z. Wang, and J. Li, "Boundary-enhanced co-training for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 574–19 584.

[27] Z. Peng, G. Wang, L. Xie, D. Jiang, W. Shen, and Q. Tian, "Usage: A unified seed area generation paradigm for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 624–634.

[28] Y. Lin, M. Chen, W. Wang, B. Wu, K. Li, B. Lin, H. Liu, and X. He, "Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 305–15 314.

[29] J. Pan, P. Zhu, K. Zhang, B. Cao, Y. Wang, D. Zhang, J. Han, and Q. Hu, "Learning self-supervised low-rank network for single-stage weakly and semi-supervised semantic segmentation," *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1181–1195, 2022.

[30] Y. Wu, X. Ye, K. Yang, J. Li, and X. Li, "Dupl: Dual student with trustworthy progressive learning for robust weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3534–3543.

[31] R. Xu, C. Wang, J. Sun, S. Xu, W. Meng, and X. Zhang, "Self correspondence distillation for end-to-end weakly-supervised semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3045–3053.

[32] L. Ru, H. Zheng, Y. Zhan, and B. Du, "Token contrast for weakly-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3093–3102.

[33] Z. Yang, Y. Meng, K. Fu, S. Wang, and Z. Song, "More: Class patch attention needs regularization for weakly supervised semantic segmentation," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[34] Z. Yang, K. Fu, M. Duan, L. Qu, S. Wang, and Z. Song, "Separate and conquer: Decoupling co-occurrence via decomposition and representation for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3606–3615.

[35] Z. Yang, Y. Meng, K. Fu, F. Tang, S. Wang, and Z. Song, "Exploring clip's dense knowledge for weakly supervised semantic segmentation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 223–20 232.

[36] Z. Xu, F. Tang, Z. Chen, Y. Su, Z. Zhao, G. Zhang, J. Su, and Z. Ge, "Toward modality gap: Vision prototype learning for weakly-supervised semantic segmentation with clip," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 9, 2025, pp. 9023–9031.

[37] S. Jang, J. Yun, J. Kwon, E. Lee, and Y. Kim, "Dial: Dense image-text alignment for weakly supervised semantic segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 248–266.