Learning Beyond Vision: Vision-Language Distillation and Edge-Aware Mix Diffusion in Semi-Supervised Semantic Segmentation

Rui Yang[1], Yunfei Bai[2], Yuehua Liu[1], Xiaomao Li[2], Shaorong Xie[1]*

[1]School of Computer Engineering and Science, Shanghai University
[2]School of Mechatronic Engineering and Automation, Shanghai University
99 Shangda Road, Dachang Town
Shanghai, 200444 China
srxie@shu.edu.cn

Abstract

In semi-supervised semantic segmentation (SSSS), segmentation performance is heavily constrained by the quality of pseudo labels. However, prevalent pseudo-label optimization approaches rely on the model's internal self-correction. When the model fails to recognize or adequately represent certain classes, this self-enhancement mechanism amplifies initial mistakes, ultimately leading to poor semantic or spatial consistency. To address this limitation, we propose ViLaDiff to enhance pseudo-label quality. Specifically, ViLaDiff employs a prompt-guided image captioning task to generate descriptive text for each input image. This represents an early attempt to integrate more flexible vision-language modeling into SSSS. We further design a vision-language fusion module to enhance semantic consistency through cross-modal interaction and dual-path knowledge distillation, ensuring coherent alignment between textual semantics and visual representations. Additionally, while language provides high-level semantic guidance, it is inherently limited in expressing fine-grained spatial structures. Therefore, we propose an edge-aware mixed-noise diffusion process. It simulates feature-level uncertainty through Gaussian perturbations and introduces class-flipping noise into the masks to model misclassification errors. A higher flipping probability is applied along mask edges, enabling boundary-aware refinement during denoising. Extensive experiments on public benchmarks validate that ViLaDiff significantly improves pseudo-label quality and segmentation performance.

Code — https://github.com/836469383/ViLaDiff.git

1 Introduction

Semi-supervised semantic segmentation (SSSS) has emerged as a promising solution to reduce the reliance on expensive pixel-level annotations (Liu et al. 2022; Sun et al. 2023a). Recent approaches typically adopt a teacher-student framework, where the teacher provides high-confidence pseudo labels to guide the student toward perturbation-invariant predictions through consistency regularization (Ouali, Hudelot, and Tami 2020; Yang et al. 2023a; Wang et al. 2024b). In this paradigm, the overall performance remains fundamentally limited by the quality of the pseudo labels (Liu et al. 2022; Yang et al. 2025).

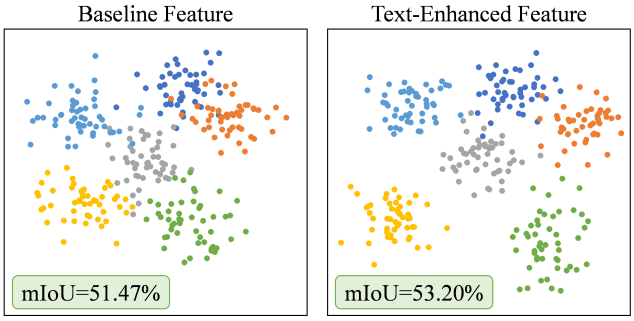

Figure 1: Feature distribution visualization based on vision with and without textual embedding. Feature embedding distributions of selected categories from the ADE20K dataset (Zhou et al. 2017) using t-SNE. RADIOv2.5 (Heinrich et al. 2025) serves as the baseline.

To improve the quality of pseudo labels, recent methods are broadly categorized into three directions: (i) Confidence-based pseudo-label filtering and mining (Wang et al. 2022; Ma et al. 2023). This line focuses on estimating pseudo-label uncertainty, selectively utilizing high-confidence predictions, and extracting informative signals based on pseudo-label categories. (ii) Structural consistency modeling (Xu et al. 2022; Mai et al. 2024). These methods leverage spatial or semantic regularities to refine pseudo labels. CISC-R (Wu et al. 2023a) aligns unlabeled images with similar labeled references to construct pixel-level correction maps. (iii) Multi-view pseudo-label reconstruction and fusion (Li et al. 2023; Hu, Jiang, and Schiele 2024). These approaches enhance pseudo-label quality by enforcing consistent predictions from multiple views. AllSpark (Wang et al. 2024a) employs cross-attention between labeled and unlabeled features to reconstruct aligned semantic representations. While effective, these methods primarily rely on information extracted from the images, such as confidence scores, structural cues, or multi-view consistency. They are limited to self-enhancement within the visual domain. Consequently, when the model fails to recognize or adequately represent certain categories, the self-enhancement mechanism amplifies initial mistakes, ultimately leading to poor semantic or spatial consistency.

*Corresponding author
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

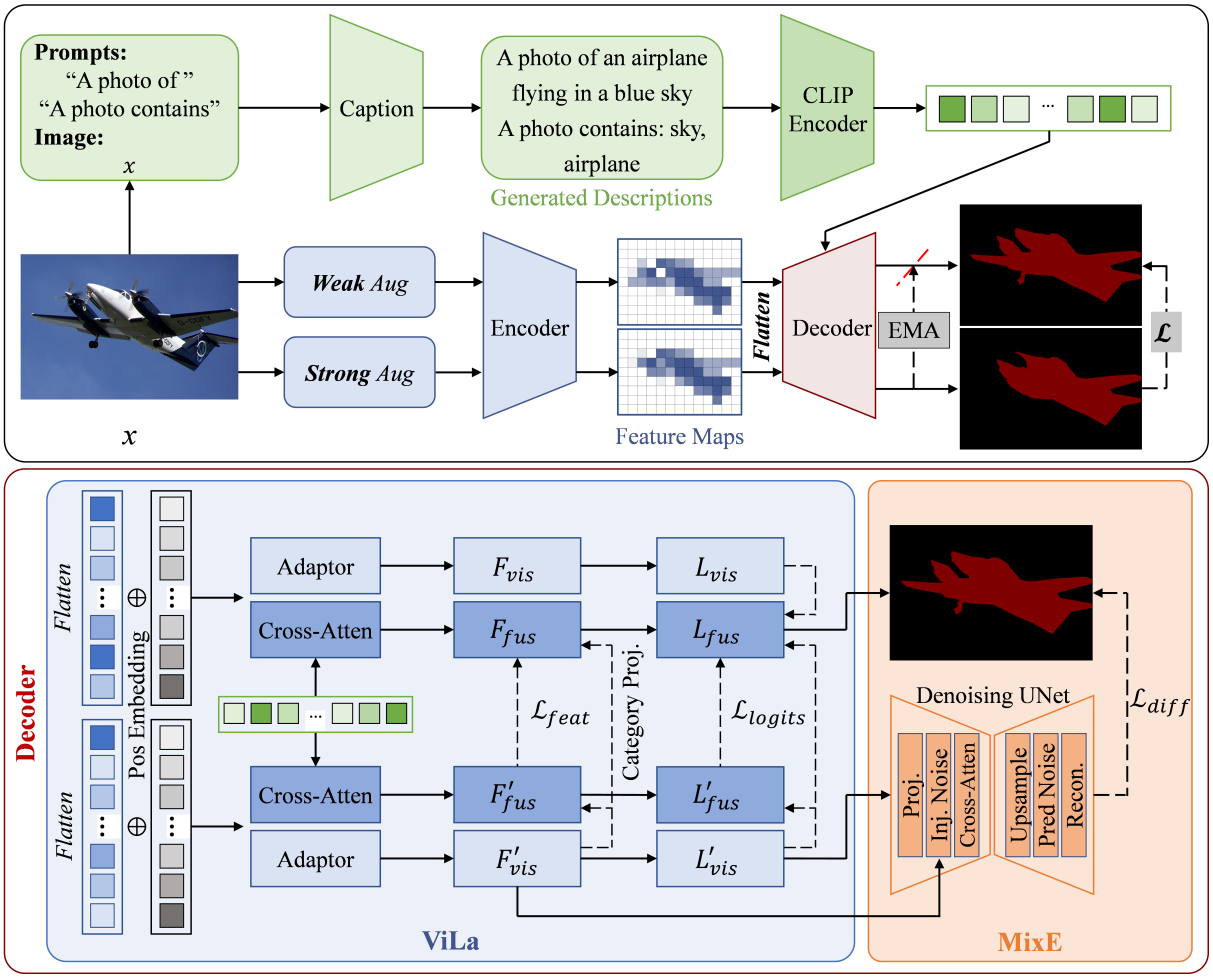

Figure 2: Overall framework pipeline. The framework comprises a text branch (green) and an image branch (blue). In the text branch, prompt-guided captioning generates descriptive text from the input image, which is then embedded using the CLIP text encoder. In the image branch, the input image is augmented and passed through an encoder to extract visual features. These features are fused with text embeddings and jointly fed into a decoder. The decoder (bottom box) incorporates two key components: ViLa and MixE. In ViLa, intermediate features are first flattened and enriched with positional encoding. They are then processed by a dual-branch structure, which extracts visual and fused features separately. Finally, distillation is applied at both the feature and logits levels. In MixE, pseudo-labels in the later training stages are refined by the denoising module. The abbreviations Proj., Inj., and Recon. in the decoder correspond to Projection, Inject, and Reconstruct, respectively.

To enhance semantic discrimination, we introduce natural language as an additional source of supervision. Textual descriptions provide explicit, human-defined semantics that directly encode category-level knowledge. These descriptions reflect intentional human cognition and naturally align with label-relevant concepts. Unlike visual signals that require complex hierarchical modeling, textual inputs serve as controllable semantic priors. They provide direct guidance for representation learning, helping the model discover and refine underrepresented classes. Motivated by this observation, we introduce natural language to guide semantic modeling for unlabeled images. Specifically, we generate image descriptions via a prompt-driven captioning task (Ghandi, Pourreza, and Mahyar 2023). To assess the effectiveness of the textual information in enhancing segmentation performance, we conduct validation experiments (see Section 3.2 for details). As shown in Figure 1, the feature distribution with text embedding exhibits more compact and well-separated clusters, demonstrating the strong semantic expressiveness and the effectiveness of the generated descriptions. Building on this insight, we propose **ViLa**, a **Vi**sion-**La**nguage Dual Distillation Module. ViLa employs a cross-modal attention mechanism to fuse image features with textual descriptions. To fully leverage semantic cues from language while avoiding over-reliance on textual inputs, ViLa adopts a dual-stream architecture comprising a fusion stream and a visual stream. Knowledge is transferred from the fusion stream to the visual stream through a dual-path distillation strategy at both feature and prediction levels, maintaining discriminative and consistent representations even in the

absence of language guidance.

The integration of natural language effectively enhances category-level semantic alignment. However, it inherently lacks the spatial information required to describe object shapes, sizes, and locations, limiting its capacity to support fine-grained boundary delineation. To enhance the spatial accuracy of pseudo labels, we first review prior methods for enforcing spatial consistency, including geometric transformation invariance (Yun et al. 2019), pixel correlation modeling (Mai et al. 2024), and boundary refinement (Guo et al. 2022). These methods largely rely on local constraints or heuristic assumptions, making them less capable of recovering severely corrupted structures. In this work, we introduce a diffusion-based model, where the iterative refinement process mirrors the decoder's progressive reconstruction of semantic structures. Through progressive denoising of corrupted inputs, it enables effective reconstruction of intricate spatial patterns and fine boundary details, addressing the shortcomings of prior heuristic-based approaches.

Although most existing diffusion-based approaches operate in continuous space, we propose to perform the denoising process directly in the discrete label space. This strategy is inspired by the observation that one-hot masks inherently encode sharp semantic boundaries and explicit category information. By modeling in discrete space, the model enables more structured reasoning over class-level transitions and spatial layouts, providing a principled solution to the challenges of pseudo-label refinement. To this end, we propose the **E**dge-aware **Mix** Diffusion (**MixE**) module, which introduces a hybrid noise injection strategy. Specifically, **category-level flipping** simulates label error and boundary ambiguity by discrete noise in mask space, while **feature-space Gaussian noise** introduces controlled perturbations that simulate realistic uncertainty and learn to produce stable predictions. This design guides the model to implicitly correct noisy pseudo labels through reconstructing semantic content. Furthermore, we introduce an edge-aware noise scheduling scheme that dynamically increases the perturbation intensity near predicted boundaries. This explicitly guides the model to focus on boundary refinement during denoising, improving the boundary accuracy of pseudo labels. The complete framework is illustrated in Figure 2.

In summary, the contributions are as follows:

- We propose **ViLaDiff**, a novel pseudo-label refinement framework that integrates vision-language fusion with diffusion modeling for semi-supervised semantic segmentation.
- We design the **ViLa** module, which performs deep image-text fusion via cross-modal attention and employs a dual-path distillation scheme to enhance representation quality and semantic consistency.
- We develop the **MixE** module, a hybrid diffusion process that integrates label flipping and feature-space noise perturbations, guided by an edge-aware scheduling mechanism to enhance spatial consistency.

2 Related Work

2.1 Semi-Supervised Semantic Segmentation

SSSS seeks to achieve accurate pixel-level predictions using limited annotated samples and large-scale unlabeled images (Fan et al. 2022; Wang et al. 2023; Zhao et al. 2023; Huang et al. 2023). This paradigm effectively alleviates the substantial manual cost associated with dense image labeling. Predominant strategies in SSSS include pseudo-labeling (Zhu et al. 2021; Yang et al. 2022; Feng et al. 2022), consistency regularization (Xie et al. 2020; Sohn et al. 2020; Sun et al. 2023a; Mai et al. 2024), and contrastive learning (Alonso et al. 2021; Zhang et al. 2022; Xie et al. 2024). Recent works increasingly integrate these approaches, forming a unified framework that leverages their complementary strengths. This trend reflects the core challenges of SSSS, which require reliable supervision, structural consistency, and discriminative representations. Pseudo-labeling offers explicit category-level supervision and serves as a foundational mechanism in semi-supervised learning. Consistency regularization improves prediction stability and model robustness by enforcing invariance under input perturbations. Contrastive learning facilitates intra-class compactness and inter-class separation, enhancing semantic discrimination. However, contrastive learning is constrained by the difficulty of constructing positive-negative pairs and high computation costs. In this work, we focus on pseudo-labeling and consistency regularization, which jointly improve pseudo-label quality and segmentation performance.

2.2 Vision-Language Model in Segmentation

Vision-language models offer high-level semantic priors about class relationships and co-occurrence patterns, enabling more generalizable multimodal segmentation frameworks (Radford et al. 2021; Yang et al. 2023b). Current approaches in this area focus primarily on open-vocabulary segmentation (Xu et al. 2023), text-guided segmentation (Marcos-Manchón et al. 2024), and zero-shot segmentation (Wu et al. 2023b). These methods typically rely on aligning textual descriptions with visual features to enhance semantic understanding. Despite the promising results, existing vision-language segmentation methods face several limitations. First, template or static textual inputs limit their scalability in large-scale semantic modeling scenes. Second, most widely used semantic segmentation datasets lack paired textual descriptions, restricting the applicability and evaluation of such methods on public benchmarks. To address these limitations, we introduce an image captioning task to generate descriptions of input images, providing high-level semantic information.

2.3 Diffusion Models in Segmentation

Diffusion models have recently garnered attention in dense prediction tasks due to their ability to capture structure and uncertainty. The iterative generation process naturally aligns with the decoder in semantic segmentation, where structured outputs are gradually reconstructed from latent representations. Existing diffusion-based segmentation methods are

broadly classified into two groups. The first leverages natural language prompts to generate class-level embeddings, which are then matched with visual features to guide generative segmentation (Barsellotti et al. 2024). The second directly employs textual descriptions to generate and localize semantically relevant content, enabling open-vocabulary, zero-shot, and cross-modal segmentation tasks (Liang et al. 2023b). Although effective in generalization, these methods operate only in continuous domains and do not explicitly model the discrete and structured nature of segmentation masks. This makes it challenging to capture semantic boundaries and address pseudo-label ambiguities, especially in complex or ambiguous regions. To address this limitation, we propose the MixE module, which injects hybrid noise at each diffusion timestep and incorporates a spatially adaptive, edge-guided noise scheduling strategy. This design enables structure-aware learning and enhances boundary reconstruction, thereby improving the quality of pseudo-labels under limited supervision.

3 Method

This section begins with an outline of the proposed method, followed by an introduction to a captioning task for generating descriptions. Building upon this, ViLa and MixE are designed to improve the quality of pseudo labels.

3.1 ViLaDiff

In this work, we first introduce natural language as auxiliary information, leveraging its high abstraction to model semantic consistency between object categories. Validation experiments show that natural language is effective in describing high-level semantic concepts. Nonetheless, it inherently struggles to convey fine-grained spatial structures, textures, and boundary details. As a result, semantic enhancement alone is insufficient to recover spatial details lost during encoder downsampling. To compensate for this limitation, we design the MixE module, which conditions the diffusion process on visual features and logits to refine the predicted mask. By injecting hybrid perturbation noise and using an edge-aware kernel, it enhances boundary awareness and structural details.

3.2 Caption Generation and Effectiveness Verification

Captioning Task. Image captioning bridges vision and language by generating natural language descriptions that match the semantic content of input images. In this study, we employ the BLIP-2 model, which uses a prompt-based mechanism to enhance the expressiveness and controllability of generated descriptions. To obtain both global semantic context and local category-specific information, we adopt two prompts: "A photo of" and "A photo contains". The caption generation is conducted offline.

Experimental Validation of Textual Guidance. To evaluate the effectiveness of caption-generated text in semantic segmentation, we conduct a validation experiment using the advanced architecture RADIOv2.5. Cross-attention is computed between image features and text embeddings, and the resulting attention maps are directly applied for mask prediction. As shown in Figure 1, incorporating text embeddings leads to clear performance gains on the ADE20K dataset, even with a state-of-the-art backbone. This demonstrates that caption-derived text provides complementary semantic cues beneficial for segmentation. Motivated by this observation, we introduce an adaptive vision–language feature fusion module, described in the next section.

3.3 ViLa

Motivated by the validation results, we propose ViLa, a dual-path distillation module for vision-language fusion built upon a teacher-student framework. It comprises a teacher and a student branch, each with a visual and a fusion stream for processing unimodal and cross-modal features, respectively. In the student model (bottom branch), strong input perturbations and partially shared architectures lead to weak feature representations and overly similar outputs. Therefore, we introduce an additional distillation path from the teacher to the student ($F'_{vis} \rightarrow F_{fus}$ and $L'_{vis} \rightarrow L_{fus}$), improving the student's stability and representation quality. When the fusion stream performs better, dual-path distillation is applied to guide the visual stream by aligning its feature representations and prediction logits. We employ a combined loss of mean squared error and cosine similarity to ensure consistency in feature magnitudes and semantic directions. The formulations for feature and prediction logits distillation are as follows:

$$\mathcal{L}_{feat} = \|F_t - F_s\|_2^2 + \frac{\lambda}{B}\sum_{b=1}^{B}\left(1 - \cos\left(f_t^{(b)}, f_s^{(b)}\right)\right) \quad (1)$$

where F_t and F_s denote the global feature representations extracted from the teacher and student branches, respectively. The first term computes the squared L2 distance. The second calculates the average cosine distance across all B samples, where $f^{(b)}$ is the normalized feature of the b-th sample.

$$\mathcal{L}_{logits} = T^2 \cdot \mathrm{KL}\left(\mathcal{S}\left(\frac{L_t}{T}\right) \| \mathcal{S}\left(\frac{L_s}{T}\right)\right) \quad (2)$$

where logits of the teacher and student models are denoted as $L_t, L_s \in \mathbb{R}^C$, respectively. The $\mathcal{S}(\cdot)$ means the function of $softmax$. The temperature parameter T controls the smoothness of the softmax distribution. The Kullback–Leibler (KL) divergence measures the discrepancy between teacher and student predictions. The term T^2 serves to balance gradient scaling, following (Hinton, Vinyals, and Dean 2015).

3.4 MixE

After introducing ViLa, we observed a performance improvement in semantic segmentation. Nevertheless, textual guidance struggles to provide fine-grained semantic boundary information. This limitation becomes pronounced after the upsampling phase, where spatial resolution is restored to generate segmentation masks. To address this, we introduce a diffusion model, which produces high-quality outputs

through a stepwise denoising process (Ho, Jain, and Abbeel 2020; Austin et al. 2021). Their capacity to model uncertainty and structure makes them naturally suited for segmentation mask generation.

Preliminary. The diffusion process consists of a forward process, which progressively adds noise to the data, and a reverse process that denoises it to reconstruct the original structure. For efficient training, the intermediate noisy state x_t is sampled directly from clean data x_0 via the following noise distribution:

$$q(x_t \mid x_0)_G = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} \cdot x_0, (1 - \bar{\alpha}_t) \cdot I) \quad (3)$$

where x_0 denotes the original input and x_t represents the noisy input at the timestep t. $x_t = \sqrt{\bar{\alpha}_t} \cdot x_0 + \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$, ϵ is standard Gaussian noise. The corruption process samples x_t independently from the distribution. The noise level at each timestep is controlled by a scalar $\alpha_t \in (0, 1)$, and the cumulative product is defined as $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

The objective of the reverse denoising process is to train a neural network ϵ_θ to predict the true noise ϵ added during the forward diffusion process, defined as:

$$\mathcal{L}_d = \mathbb{E}_{x_0, \epsilon, t}[||\epsilon - \epsilon_\theta(x_t, t, cond)||^2] \quad (4)$$

where $\mathbb{E}$ is the mathematical expectation.

MixE. In this work, we design a class-flipping-based pseudo diffusion mechanism to simulate discrete misclassification in pseudo-labels. The conditional distribution from the initial label state x_0 to a corrupted label at timestep t, denoted as x_t, is formulated as:

$$q(x_t \mid x_0)_C = \prod_i Cls\left(x_t^{(i)} \mid \pi_t^{(i)}(x_0^{(i)})\right) \quad (5)$$

where $x_t^{(i)}$ denotes the category of the i-th pixel at timestep t, and $\pi_t^{(i)}(x_0^{(i)}) := T_t^i[:, x_0^{(i)}]$ represents a categorical distribution derived from a class transition matrix T_t^i. Each $T_t^i(c \mid k)$ models the probability of flipping from the original class k to class c at timestep t, defined as:

$$T_t^i(c \mid k) = \begin{cases} 1 - \beta_t^{(i)}, & \text{if } c = k \\ \frac{\beta_t^{(i)}}{C-1}, & \text{if } c \neq k \end{cases} \quad (6)$$

where $\beta_t = 0.1$ controls the corruption strength and C represents the total number of categories. This process preserves the original label with probability $1 - \beta_t$, while uniformly flipping it to any other $C - 1$ classes with probability β_t. To further improve boundary sensitivity, a Laplacian kernel is used to detect mask edges, where the flipping probability is increased by 0.1.

MixE combines class-flipping and Gaussian noise injection to model two types of pseudo-label imperfection: category misclassification and spatial uncertainty, respectively. For each pixel i, the perturbation type is determined by its pseudo-label confidence score $\gamma^{(i)} = max_c x_0^{(i)}(c)$, where $x_0^{(i)} \in \mathbb{R}^C$ denotes the soft pseudo label. A threshold τ is exploited to separate high- and low-confidence regions. Specifically, we defined the forward corruption process as:

$$q(x_t^{(i)} \mid x_0^{(i)}) = \begin{cases} q_G(x_t^{(i)}, \mid x_0^{(i)}), & \text{if } \gamma^{(i)} > \tau \\ q_C(x_t^{(i)}, \mid \hat{y}_0^{(i)}), & \text{if } \gamma^{(i)} \leq \tau \end{cases} \quad (7)$$

where $\hat{y}_0^{(i)} = argmax_c x_0^{(i)}$ is the predicted class. q_C denotes a categorical diffusion with the transition matrix T_t to simulate class-level perturbations. q_G applies Gaussian noise to feature embeddings to model uncertainty in ambiguous regions. The confidence threshold τ increases linearly from 0 to 0.8 during the first half of training and then remains fixed.

During the reverse denoising, the model reconstructs the clean segmentation masks by estimating the posterior distribution over semantic labels at each pixel. A conditional denoising model f_θ is used to approximate the pixel-wise categorical distribution for the previous state $x_{t-1}^{(i)}$, conditioned on the current noisy label $x_t^{(i)}$ and auxiliary input $cond$. The predicted distribution is defined as:

$$p_\theta(x_{t-1}^{(i)} \mid x_t^{(i)}, cond) = \mathcal{S}(f_\theta(x_t^{(i)}, t, cond)) \quad (8)$$

where $cond$ includes both noisy visual features and classification logits.

3.5 Loss Functions

The overall training objective comprises two parts, corresponding to labeled or unlabeled data:

$$\mathcal{L} = \mathcal{L}_{sup} + \mathcal{L}_{uns} \quad (9)$$

For labeled samples, we adopt the standard multi-class cross-entropy loss:

$$\mathcal{L}_{sup} = -\frac{1}{N} \sum_{i=1}^{N} logp(i, y_i) \quad (10)$$

For unlabeled data, the loss consists of two components: a bi-branch consistency loss and a diffusion-based refinement loss:

$$\mathcal{L}_{uns} = \mathcal{L}_{cons} + \mu \cdot \mathcal{L}_{diff} \quad (11)$$

The weight vector μ is set to [1, 0.5, 0.5]. The consistency loss L_{cons} encourages alignment between the student and teacher predictions across both the pure visual and fused branches. Specifically:

$$\mathcal{L}_{cons} = \mathcal{L}_{feat}(F_{fus}, F_{vis}) + \mathcal{L}_{logits}(L_{fus}, L_{vis}) \quad (12)$$

where F_{fus}, L_{fus} and F_{vis}, L_{vis} denote the feature maps and classification logits from the fusion and visual streams, respectively. The diffusion module is supervised via:

$$\mathcal{L}_{diff} = \mathcal{L}_{ce} + \mathcal{L}_{edge} + \mathcal{L}_d \quad (13)$$

$\mathcal{L}_{ce}$ and $\mathcal{L}_{edge}$ are cross-entropy-based losses for region accuracy and boundary precision, respectively, while $\mathcal{L}_d$ supervises feature denoising.

4 Experiments

4.1 Experimental Settings

Datasets. To comprehensively evaluate the effectiveness of ViLaDiff, we conduct experiments on three standard semi-supervised semantic segmentation benchmarks: PASCAL VOC 2012 (Everingham et al. 2015), Cityscapes (Cordts et al. 2016), and COCO-Stuff (Caesar, Uijlings, and Ferrari

PASCAL HQ	1/16 (92)	1/8 (183)	1/4 (366)	1/2 (732)	Full (1464)
SupBaseline	50.7	63.6	70.8	75.6	77.2
PseudoSeg (Zou et al. 2020)	57.6	65.5	69.1	72.4	-
CPS (Chen et al. 2021)	64.1	67.4	71.7	75.9	-
U^2PL (Wang et al. 2022)	68.0	69.2	73.7	76.2	79.5
PS-MT (Liu et al. 2022)	65.8	69.6	76.6	78.4	80.0
UniMatch (Yang et al. 2023a)	75.2	77.2	78.8	79.9	81.2
DAW (Sun et al. 2023b)	74.8	77.4	79.5	80.6	81.5
CorrMatch (Sun et al. 2023a)	76.4	78.5	79.4	80.6	81.8
DDFP (Wang et al. 2024b)	75.0	78.0	79.5	81.2	82.0
AllSpark (Wang et al. 2024a)	76.1	78.4	79.8	80.8	82.1
ViLaDiff	**78.7**	**80.1**	**81.5**	**82.4**	**83.9**

Table 1: Comparison with SOTA methods on the PASCAL HQ dataset. The table presents mIoU (%) under various protocols.

PASCAL AUG	1/16 (662)	1/8 (1323)	1/4 (2646)	1/2 (5291)
SupBaseline	72.0	73.1	76.7	77.6
CPS (Chen et al. 2021)	72.2	75.8	77.6	78.6
PS-MT (Liu et al. 2022)	75.5	78.2	78.7	79.8
UniMatch (Yang et al. 2023a)	78.1	78.4	79.2	-
DAW (Sun et al. 2023b)	78.5	78.9	79.6	-
CorrMatch (Sun et al. 2023a)	78.4	79.3	79.6	-
DDFP (Wang et al. 2024b)	78.3	78.9	79.8	80.9
AllSpark (Wang et al. 2024a)	78.3	80.0	80.4	81.1
ViLaDiff (Ours)	**79.4**	**80.2**	**82.1**	**82.7**
U^2PL† (Wang et al. 2022)	68.0	69.2	73.7	76.2
UniMatch† (Yang et al. 2023a)	80.9	81.9	80.4	-
AllSpark† (Wang et al. 2024a)	81.6	82.0	80.9	81.1
ViLaDiff (Ours)†	**79.7**	**80.3**	**81.6**	**82.9**

Table 2: Comparison with SOTA methods on the PASCAL AUG dataset. † indicates same splits as U^2PL.

Cityscapes	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)
SupBaseline	65.7	70.4	73.8	78.1
CPS (Chen et al. 2021)	69.8	74.3	74.6	76.8
UniMatch (Yang et al. 2023a)	76.6	77.9	79.2	79.5
Swicth (Na et al. 2023)	76.8	78.4	79.4	80.5
DAW (Sun et al. 2023b)	76.6	78.4	79.8	80.6
LogicDiag (Liang et al. 2023a)	76.8	78.9	80.2	81.3
CorrMatch (Sun et al. 2023a)	77.3	78.5	79.4	80.4
SemiVL† (Hoyer et al. 2024)	77.9	79.4	80.3	80.6
DDFP (Wang et al. 2024b)	77.1	78.2	79.9	80.8
AllSpark (Wang et al. 2024a)	78.3	79.2	80.6	81.4
ViLaDiff (Ours)	**80.5**	**81.2**	**82.4**	**83.3**

Table 3: Comparison with SOTA methods on the Cityscapes dataset.

2018). These datasets differ in scene complexity and category granularity, ensuring a thorough and authoritative assessment. **PASCAL VOC 2012** consists of 21 object-centric classes in natural scenes. To enrich data diversity, additional samples are incorporated from the Semantic Boundaries Dataset (SBD), forming an augmented training set. The dataset presents challenges such as appearance variation, occlusion, and pose changes, making it a widely adopted benchmark. **Cityscapes** focuses on urban street scenes and provides 5,000 high-resolution images with fine-grained pixel annotations. It includes 2,975 training, 500 validation, and 1,525 test images, covering 19 driving-related categories. Its fine detail and high resolution make it a strong benchmark for evaluating boundary sensitivity and small object segmentation. **COCO-Stuff** offers 118,000 training and 5,000 validation images across 81 categories. Known for its complex scenes and dense semantics, COCO is a standard benchmark for testing model scalability and generalization in large-scale, multi-class segmentation tasks.

Implementation Details. To ensure fair comparison, we adopt the standard ViT-B/16 model as the feature extractor. Given the complexity of gradient dynamics, all experiments are optimized using AdamW with a learning rate of 0.001, default β values, and a batch size of 4. The teacher model is updated using an exponential moving average (EMA) with a momentum of 0.99. The forward Gaussian diffusion schedule follows the setting of DDPM (Ho, Jain, and Abbeel 2020). All models are implemented in PyTorch and trained using mixed-precision on 2×RTX 3090Ti GPUs with distributed data parallelism.

4.2 Comparison with SOTA

Following previous protocols, we conduct a comprehensive comparison between ViLaDiff and SOTA methods.

HQ PASCAL VOC 2012. Table 1 summarizes the results on the HQ PASCAL VOC 2012 dataset. Compared to the fully supervised baseline (first row), our method achieves performance gains of 28.0%, 16.5%, 10.7%, 6.8%, and 6.7% under data splits 1/16, 1/8, 1/4, 1/2, and Full. Furthermore, ViLaDiff outperforms advanced semi-supervised segmentation approaches such as AllSpark by 1.8%, demonstrating its strong capability in enhancing pseudo-label quality for supervision.

AUG PASCAL VOC 2012. Table 2 presents the performance of our method on the Augmented PASCAL VOC dataset. Under low-data regimes, our approach consistently yields stable improvements, indicating that textual features provide clear semantic guidance and enhance discriminative ability when annotations are limited. Moreover, under the "Full" supervision setting, ViLaDiff achieves the best performance. This suggests that even when texts yield diminishing returns with sufficient data, MixE still further improves pseudo-label quality through boundary refinement and structural consistency modeling.

Cityscapes. As shown in Table 3, our method achieves the best performance across all data split settings. The consistent improvements underscore its strength in capturing fine boundary details and accurately segmenting small objects in high-resolution scenes.

COCO-Stuff. Table 4 presents a comparison between our method and current mainstream approaches on the COCO-Stuff dataset. Across various data partition settings, our method consistently achieves superior segmentation performance, demonstrating its enhanced discriminative capability in handling complex scenes and semantically dense images.

COCO-Stuff	1/512 (232)	1/256 (463)	1/128 (925)	1/64 (1849)
SupBaseline	19.7	26.8	35.9	41.2
PseudoSeg	29.8	37.1	39.1	41.8
UniMatch	31.9	38.9	44.4	48.2
LogicDiag	33.1	40.3	45.4	48.8
AllSpark	34.1	41.7	45.5	49.6
ViLaDiff (Ours)	**36.5**	**43.9**	**47.6**	**50.7**

Table 4: Comparison with SOTA methods on the COCO-Stuff dataset.

Baseline	ViLa		MixE			1/8 (183)	1/2 (732)
	$\mathcal{L}_{feat}^{(t)}$	$\mathcal{L}_{logits}^{(t)}$	q_G	q_C	$\mathcal{L}_{edge}$		
✓						72.2	75.0
✓	✓					74.5	77.3
✓		✓				73.7	76.1
✓	✓	✓				75.3	77.4
✓	✓	✓	✓			77.8	79.6
✓	✓	✓		✓		78.1	80.5
✓	✓	✓	✓	✓		79.6	81.8
✓	✓	✓	✓	✓	✓	**80.1**	**82.4**

Table 5: Ablation study of proposed modules on PASCAL VOC 2012 HQ using a ViT-based SemiBaseline is used for evaluation.

4.3 Ablation Studies

Effectiveness of ViLaDiff Modules. To clarify the contribution of each component to the overall performance, we conducted ablation studies. Notably, the ablation of the ViLa is performed only on the teacher model, since its impact mainly arises from distillation on the teacher side. The student model's distillation ($F'_{vis} \rightarrow F'_{fus}$ and $L'_{vis} \rightarrow L'_{fus}$) is small and primarily serves to reduce dependence on textual features during inference, ensuring input flexibility. The results indicate that each module demonstrates a distinct and indispensable contribution to the overall performance.

Balancing Coefficient of Distillation Loss. The feature distillation loss $\mathcal{L}_{feat}$ in ViLa combines MSE and cosine distance to enhance the model's sensitivity to differences in magnitude and direction of intermediate features. Due to differing gradient sensitivities, assigning equal weights causes directional loss to dominate. This results in training instability and slow convergence. To address this, we incorporate a balancing coefficient λ to regulate the relative contribution of each loss component. Through systematic ablation experiments, we evaluated and selected the optimal λ to ensure stable and effective distillation signals. As shown in Table 6, the model performance exhibits consistent trends across various datasets and data splits. The best performance is achieved at $\lambda = 0.5$, indicating an optimal balance between magnitude and directional contributions. Conversely, larger or smaller λ values amplify or suppress the component, negatively affecting overall performance.

λ	PASCAL VOC		COCO-Stuff	
	1/16 (92)	1/2 (732)	1/256 (463)	1/128 (925)
0.1	77.5	81.2	43.0	46.5
0.3	78.1	81.8	43.5	47.2
0.5	**78.7**	**82.4**	**43.9**	**47.6**
0.8	78.5	82.1	43.1	46.9
1.0	77.4	80.6	42.7	46.3

Table 6: Ablation Study on the Feature Distillation Loss Weight.

Timesteps	PASCAL VOC		COCO-Stuff	
	1/16 (92)	1/2 (732)	1/256 (463)	1/128 (925)
5	74.9	77.2	40.1	43.6
10	76.3	78.5	41.6	44.9
15	76.3	78.7	41.7	45.0
20	76.4	**78.9**	41.7	**45.2**
30	**76.5**	78.8	**41.9**	45.1

Table 7: Ablation study on diffusion timesteps under various data regimes on PASCAL VOC and COCO 2017. Ablation experiments are performed based on ViT-B.

Ablation Study on Diffusion Steps. The number of diffusion steps controls the noise injection and denoising depth in diffusion models. Fewer steps facilitate easier denoising and more stable training, but limit the model's ability to capture complex structures. More steps improve generation quality but cause training instability. To balance these factors, we conduct an ablation study on diffusion steps. Experimental results in Table 7 indicate that increasing the number of denoising steps yields diminishing returns, particularly after step 10. Accordingly, we fix the denoising step to 10 for all experiments in this paper.

5 Conclusion

To enhance pseudo-label quality in SSSS, this paper proposes a pseudo-label refinement framework that integrates vision-language dual distillation (ViLa) and edge-aware mixed diffusion module (MixE). ViLa leverages text generated from a captioning model to effectively guide the model's attention to target regions, improving semantic consistency. During training, a dual-stream distillation approach is employed to achieve stable and efficient knowledge transfer. Meanwhile, MixE injects semantic noise to refine mask boundaries and structural consistency further. Extensive experiments demonstrate that ViLaDiff achieves superior performance across multiple datasets, validating its effectiveness in significantly improving pseudo-label quality.

Acknowledgments

This work was supported by the Science Fund for Creative Research Groups of the National Natural Science Foundation of China (No. 62421004).

References

Alonso, I.; Sabater, A.; Ferstl, D.; Montesano, L.; and Murillo, A. C. 2021. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8219–8228.

Austin, J.; Johnson, D. D.; Ho, J.; Tarlow, D.; and Van Den Berg, R. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34: 17981–17993.

Barsellotti, L.; Amoroso, R.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2024. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3689–3698.

Caesar, H.; Uijlings, J.; and Ferrari, V. 2018. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1209–1218.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021. Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2613–2622.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111: 98–136.

Fan, J.; Gao, B.; Jin, H.; and Jiang, L. 2022. UCC: Uncertainty guided Cross-head Cotraining for Semi-Supervised Semantic Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9937–9946.

Feng, Z.; Zhou, Q.; Gu, Q.; Tan, X.; Cheng, G.; Lu, X.; Shi, J.; and Ma, L. 2022. Dmt: Dynamic mutual training for semi-supervised learning. *Pattern Recognition*, 130: 108777.

Ghandi, T.; Pourreza, H.; and Mahyar, H. 2023. Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, 56(3): 1–39.

Guo, H.; Du, B.; Zhang, L.; and Su, X. 2022. A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183: 240–252.

Heinrich, G.; Ranzinger, M.; Yin, H.; Lu, Y.; Kautz, J.; Tao, A.; Catanzaro, B.; and Molchanov, P. 2025. Radiov2. 5: Improved baselines for agglomerative vision foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22487–22497.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hoyer, L.; Tan, D. J.; Naeem, M. F.; Van Gool, L.; and Tombari, F. 2024. SemiVL: semi-supervised semantic segmentation with vision-language guidance. In *European Conference on Computer Vision*, 257–275. Springer.

Hu, X.; Jiang, L.; and Schiele, B. 2024. Training Vision Transformers for Semi-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4007–4017.

Huang, H.; Xie, S.; Lin, L.; Tong, R.; Chen, Y.-W.; Li, Y.; Wang, H.; Huang, Y.; and Zheng, Y. 2023. SemiCVT: Semi-Supervised Convolutional Vision Transformer for Semantic Segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11340–11349.

Li, S.; He, Y.; Zhang, W.; Zhang, W.; Tan, X.; Han, J.; Ding, E.; and Wang, J. 2023. CFCG: semi-supervised semantic segmentation via cross-fusion and contour guidance supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16348–16358.

Liang, C.; Wang, W.; Miao, J.; and Yang, Y. 2023a. Logic-induced diagnostic reasoning for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16197–16208.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023b. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7061–7070.

Liu, Y.; Tian, Y.; Chen, Y.; Liu, F.; Belagiannis, V.; and Carneiro, G. 2022. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4258–4267.

Ma, J.; Wang, C.; Liu, Y.; Lin, L.; and Li, G. 2023. Enhanced soft label for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1185–1195.

Mai, H.; Sun, R.; Zhang, T.; and Wu, F. 2024. RankMatch: Exploring the better consistency regularization for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3391–3401.

Marcos-Manchón, P.; Alcover-Couso, R.; SanMiguel, J. C.; and Martínez, J. M. 2024. Open-vocabulary attention maps with token optimization for semantic segmentation in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9242–9252.

Na, J.; Ha, J.-W.; Chang, H. J.; Han, D.; and Hwang, W. 2023. Switching temporary teachers for semi-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 36: 40367–40380.

Ouali, Y.; Hudelot, C.; and Tami, M. 2020. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12674–12684.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608.

Sun, B.; Yang, Y.; Zhang, L.; Cheng, M.-M.; and Hou, Q. 2023a. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. *arXiv preprint arXiv:2306.04300*.

Sun, R.; Mai, H.; Zhang, T.; and Wu, F. 2023b. DAW: Exploring the Better Weighting Function for Semi-supervised Semantic Segmentation. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 61792–61805. Curran Associates, Inc.

Wang, H.; Zhang, Q.; Li, Y.; and Li, X. 2024a. Allspark: Reborn labeled features from unlabeled in transformer for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3627–3636.

Wang, X.; Bai, H.; Yu, L.; Zhao, Y.; and Xiao, J. 2024b. Towards the uncharted: Density-descending feature perturbation for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3303–3312.

Wang, Y.; Wang, H.; Shen, Y.; Fei, J.; Li, W.; Jin, G.; Wu, L.; Zhao, R.; and Le, X. 2022. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4248–4257.

Wang, Z.; Zhao, Z.; Xing, X.; Xu, D.; Kong, X.; and Zhou, L. 2023. Conflict-based cross-view consistency for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19585–19595.

Wu, L.; Fang, L.; He, X.; He, M.; Ma, J.; and Zhong, Z. 2023a. Querying labeled for unlabeled: Cross-image semantic consistency guided semi-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7): 8827–8844.

Wu, L.; Zhang, W.; Jiang, T.; Yang, W.; Jin, X.; and Zeng, W. 2023b. [CLS] Token is All You Need for Zero-Shot Semantic Segmentation. *arXiv preprint arXiv:2304.06212*.

Xie, H.; Wang, C.; Zhao, J.; Liu, Y.; Dan, J.; Fu, C.; and Sun, B. 2024. PRCL: Probabilistic representation contrastive learning for semi-supervised semantic segmentation. *International Journal of Computer Vision*, 132(10): 4343–4361.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; and Le, Q. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33: 6256–6268.

Xu, H.; Liu, L.; Bian, Q.; and Yang, Z. 2022. Semi-supervised semantic segmentation with prototype-based consistency regularization. *Advances in neural information processing systems*, 35: 26007–26020.

Xu, J.; Hou, J.; Zhang, Y.; Feng, R.; Wang, Y.; Qiao, Y.; and Xie, W. 2023. Learning open-vocabulary semantic segmentation models from natural language supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2935–2944.

Yang, L.; Qi, L.; Feng, L.; Zhang, W.; and Shi, Y. 2023a. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7236–7246.

Yang, L.; Zhuo, W.; Qi, L.; Shi, Y.; and Gao, Y. 2022. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4268–4277.

Yang, R.; Bai, Y.; Liu, C.; Liu, Y.; Li, X.; and Xie, S. 2025. Hybrid Architectures Ensemble Learning for pseudo-label refinement in semi-supervised segmentation. *Information Fusion*, 116: 102791.

Yang, S.; Qu, T.; Lai, X.; Tian, Z.; Peng, B.; Liu, S.; and Jia, J. 2023b. LISA++: An Improved Baseline for Reasoning Segmentation with Large Language Model. *arXiv preprint arXiv:2312.17240*.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6023–6032.

Zhang, J.; Wu, T.; Ding, C.; Zhao, H.; and Guo, G. 2022. Region-level contrastive and consistency learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:2204.13314*.

Zhao, Z.; Yang, L.; Long, S.; Pi, J.; Zhou, L.; and Wang, J. 2023. Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11350–11359.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 633–641.

Zhu, Y.; Zhang, Z.; Wu, C.; Zhang, Z.; He, T.; Zhang, H.; Manmatha, R.; Li, M.; and Smola, A. 2021. Improving semantic segmentation via efficient self-training. *IEEE transactions on pattern analysis and machine intelligence*, 46(3): 1589–1602.

Zou, Y.; Zhang, Z.; Zhang, H.; Li, C.-L.; Bian, X.; Huang, J.-B.; and Pfister, T. 2020. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*.