

SESC²: Seeking Essential Semantics in Changing Contexts for Semi-Supervised Segmentation

Haifeng Zhong, Fan Tang, *Member, IEEE*, Tong-Yee Lee, *Senior Member, IEEE*, Runyang Feng, Yixing Gao, *Member, IEEE*

Abstract—Semi-supervised semantic segmentation based on consistency regularization typically relies on contextual priors derived from limited labeled data to exploit features from unlabeled data. To this end, recent state-of-the-art methods tend to optimize unlabeled features that deviate from the labeled feature distribution by aligning them with labeled data. Although this mechanism allows the model to maximize segmentation accuracy on unlabeled data with similar distributions via context recurrence, it inadvertently misinterprets unseen context in unlabeled data as anomalies. To break this closed-loop reliance on labeled contexts, we propose a new framework, Seeking Essential Semantics in Changing Environments (SESC²), which aims to compel the model to recognize objects based on object-centric semantic cues rather than spurious environmental correlations found in labeled data. Our method comprises two key components: (1) Visual Content Alteration (VCA), a strategy that dynamically exchanges backgrounds between labeled and unlabeled instances of the same category, breaking the contextual dependency imposed by labeled data while constructing diverse contexts that facilitate object-centric feature learning; and (2) Semantic Invariance Recombination (SIR), which models high-order feature interactions between the original and VCA-altered inputs, encouraging the model to extract semantic attributes that are more robust to contextual changes. Extensive experiments on multiple benchmarks demonstrate that our SESC² consistently improves baseline performance across different data partitions, particularly in low-data regimes. For instance, under the Pascal VOC 1/16 partition setting, our method improves the performance of SWSEG by 2.07% in mIoU.

Index Terms—Semantic segmentation, semi-supervised learning, consistency regularization

I. INTRODUCTION

SEMANtic segmentation is a fundamental task in computer vision, with extensive applications in fields such as scene understanding [1], autonomous driving [2], and robotic perception [3]. However, its data-driven nature makes acquiring massive pixel-level annotated data both labor-intensive and time-consuming. Consequently, semi-supervised semantic

Haifeng Zhong, Runyang Feng, and Yixing Gao are with the School of Artificial Intelligence, Jilin University, Changchun, 130012, China. E-mail: zhonghf23@mails.jlu.edu.cn and gaoyixing@jlu.edu.cn. Fan Tang is with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100086, China. Tong-Yee Lee is with the Department of Computer Science and Information Engineering, National Cheng-Kung University, 70101, Tainan, Taiwan. (*Corresponding author: Yixing Gao*)

This work was supported by the National Natural Science Foundation of China under Grant No. W2421093, and the International Cooperation Project of Jilin Province under Grant No. 20250205079GH. This work was also supported by the National Natural Science Foundation of China under Grant No. 62572458 and the Beijing Major Science and Technology Project under No. Z251100008125009. This work was also supported by the National Science and Technology Council, Taiwan under Grant 114-2221-E-006-114-MY3.

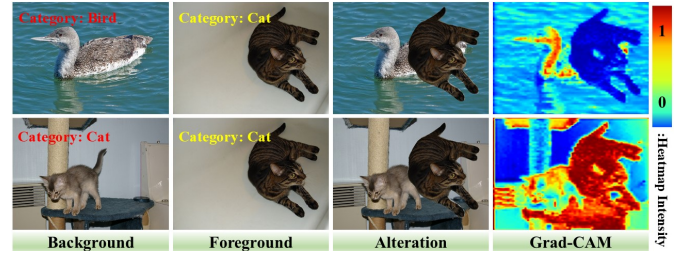


Fig. 1. The feature differences resulting from content alteration between images of the same and different class (using UniMatch and ResNet-101).

segmentation (SSSS) [4]–[6] has attracted increasing attention due to its ability to maintain high segmentation accuracy while significantly reducing annotation costs.

In the field of SSSS, pseudo-label training based on consistency regularization [4], [7] has emerged as the mainstream paradigm for leveraging unlabeled data. This paradigm typically relies on contextual priors provided by limited labeled data [5], expecting to make predictions on unlabeled data through consistency under perturbations. Consequently, their performance upper bound largely depends on whether the contextual correlations learned from a small amount of labeled data are sufficiently reliable and generalizable. However, the object-context relationships in real-world multimedia scenarios are not always conventional or stable. Particularly with the rapid proliferation of AI-generated images [8]–[10], the generated content frequently exhibits contextual inconsistency characterized by locally realistic appearance but globally imbalanced semantics (e.g., a fishing boat implausibly appearing on land). Under such application settings, if a segmentation model overly relies on the common contextual patterns established from limited labeled data, it is prone to misclassify these unconventional yet valid target regions as anomalous noise, thereby limiting its deployability in practical applications, such as AI image forensics based on semantic contextual analysis [11]. To this end, some methods [4], [12]–[15] typically employed random mixing of image regions [13] or category patches [14] between labeled data to construct diverse contextual environments for utilizing unlabeled data. However, such random content exchange ignores the semantic plausibility of the newly generated data. As illustrated in the first row of Fig. 1, the unreasonable combination of placing a cat in a pond results in counterfactual contextual semantic conflicts, thereby distracting the model’s focus from the cat.

To address the semantic confusion caused by random image mixing while strengthening the utilization of unlabeled data, recent advanced methods [5], [16]–[20] have shifted

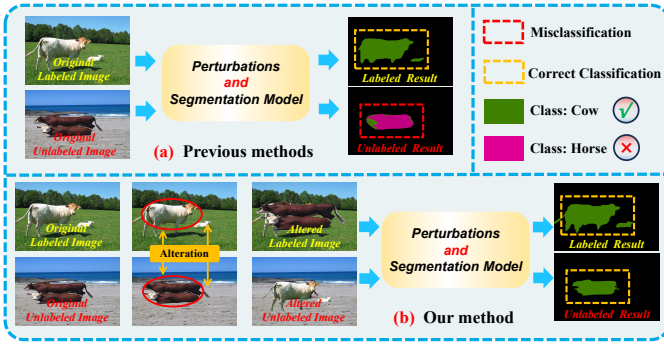


Fig. 2. Illustration of our visual context alteration strategy. Previous methods (a) rely on the contextual recurrence patterns in labeled data, which leads to misclassification when encountering cows appearing on the beach in unlabeled data. In contrast, our method (b) leverages background transformation to encourage the model to focus on the cows and to model their characteristics, thereby enabling accurate predictions in unfamiliar contextual environments by leveraging object-centric semantic cues.

towards the optimization of pseudo-labels, aiming to refine supervisory signals through feature correlations or multi-scale consistency, thereby strengthening the reliability of mining supervisory signals from unlabeled data. For instance, CorrMatch [16] and RankMatch [17] utilize the correlation or rank stability of class predictions under perturbations to filter out unreliable pseudo-labels; MLLC [19] exploits graph-based structural relationships to rectify erroneous predictions by distilling proximate patterns from semantic neighbors; whereas SWSEG [7], MPMC [18], and MGCT [20] introduce metric-based constraints or feature knowledge alignment, further purifying supervision signals by optimizing the uniformity of the feature space or integrating multi-view feature consistency.

Although the above methods [5], [16]–[20] are effective to some extent, they universally treat the contextual patterns in labeled data as the correct “gold standard.” Fundamentally, regardless of whether they utilize correlation matching (e.g., MPMC [18] and CorrMatch [16]) or employ rank stability as a filtering mechanism (e.g., SWSEG [7] and RankMatch [17]), these methods follow a verification logic: “Does this unlabeled sample align with a known pattern in the labeled data?” While this mechanism allows the model to maximize segmentation accuracy on unlabeled data with similar distributions via context recurrence, it inadvertently misinterprets valid unseen context as statistical anomalies. For example, as shown in Fig. 2, when an object in unlabeled data (e.g., a cow on a beach) violates the established correlations found in the labeled data (e.g., cows co-occurring with grass), the model tends to suppress these valuable hard samples. Consequently, instead of helping the model learn to comprehend the intrinsic features of objects in unfamiliar environments, these methods reinforce the reliance on existing contextual semantics. In this regard, object-centric representation learning (OCRL) methods [21], [22] provide us with valuable inspiration. This line of methods emphasizes that representation learning should not excessively rely on scene contexts, but should instead build stable representations predominantly around the objects themselves. However, existing OCRL methods primarily rely on object appearance similarity or spatial distribution proximity to cluster homologous content within the feature space. While such mechanisms are relatively effective in constrained

scenarios with uniform content (such as microscopy images), in complex multi-object natural scenes, relying solely on internal constraints of object similarity makes it difficult to truly decouple target features from scene contexts.

Based on the above analysis, it is natural to conceive that to fully exploit unlabeled data, it is imperative to break the contextual shackles imposed by labeled data and endow the model with the ability to capture object-related features that are less sensitive to scene variations [14], [23]. An intuitive solution appears to be the introduction of more aggressive data augmentation than previous works [13], [14] to simulate contextual environmental changes, such as mixing larger image regions or swapping patches across more categories. However, such unstructured random variations inevitably introduce unreasonable semantic combinations that distract the model’s focus. Furthermore, under context transformations driven by labeled data, how to construct robust consistency constraints to model object-centric semantic cues from complex variations remains a pressing challenge. To address these issues, we need to establish a causal pathway [24] that progresses from constructing plausible diverse context to break contextual dependencies, to learning semantic features that remain stable across diverse contexts. In light of this analysis, this paper proposes the Seeking Essential Semantics in Changing Contexts (SESC²) strategy. The innovation of SESC² lies in abandoning unstructured context transformations that compromise semantic rationality, and instead decoupling the dependency on specific environments by constructing semantically plausible dynamic contexts, thereby reformulating the semi-supervised segmentation problem as a task of learning semantic representations that are robust to changing environments.

Specifically, our SESC² consists of closely connected Visual Context Alteration (VCA) and Semantic Invariance Recombination. First, for the input image, VCA exchanges backgrounds between images with similar content (e.g., both labeled and unlabeled data contain the ‘cow’ category), allowing the same object to appear in different contexts. Through the context alteration of VCA, we can, on the one hand, alter the contextual monotony of labeled data, and on the other hand, shift the model’s focus from global image to object-centric representations under dynamically changing backgrounds. Since VCA focuses on exchanging background contexts between the same objects, it effectively avoids unreasonable semantic contexts caused by random image exchange (as shown in Fig. 1, background exchange between cats does not cause attention distraction). Next, we need to utilize consistency constraints before and after the VCA alteration to model the features of objects that are independent of context changes. To this end, we propose the Semantic Invariance Recombination (SIR). SIR goes beyond simple feature alignment; by capturing high-order feature interactions between the original view and the VCA-transformed view, it distills stable semantics from dynamic changes, thereby reconstructing semantic representations capable of penetrating environmental interference. As shown in Fig. 2(b), since our SESC² helps the model capture object-related features that are less entangled with context, even when facing unfamiliar context in unlabeled data, SESC² can enable the model to achieve accurate segmentation based

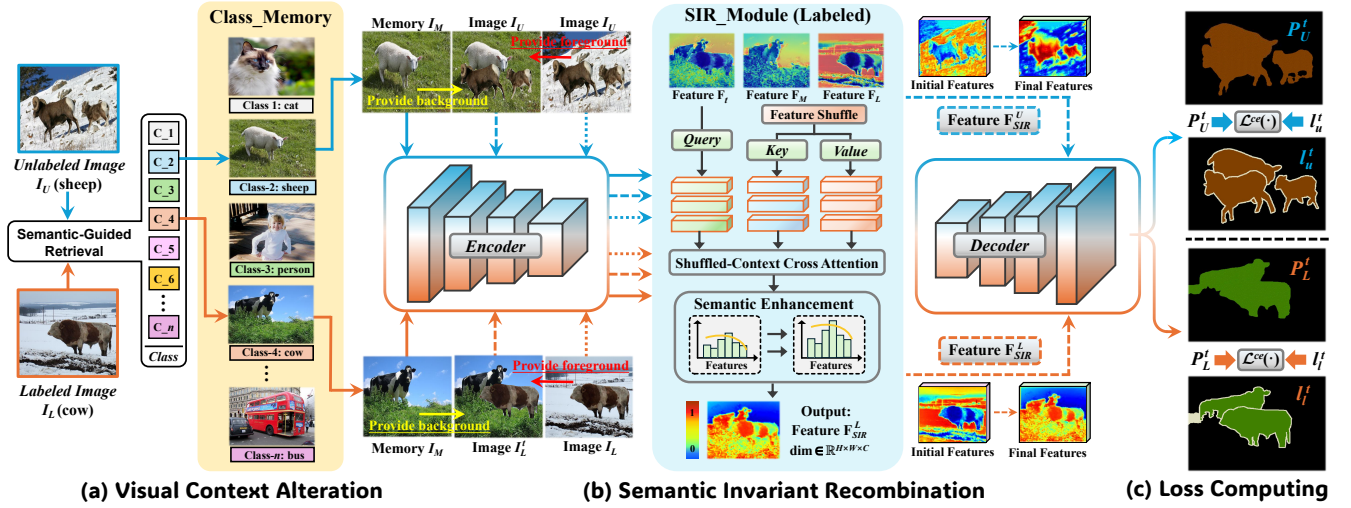


Fig. 3. Overview of our SESC². The left part depicts our method for the foreground-background transformation of labeled and unlabeled data based on **visual context alteration** (in Sec. III-B). The mid part provides a detailed description of the **semantic invariant recombination** taking labeled images as an example (in Sec. III-C). The right part shows the **details of loss computation** (in Sec. III-D) of our method.

on object-centered representations.

We conduct extensive experiments on three semi-supervised segmentation benchmarks (Pascal VOC [25], Cityscapes [26], and COCO [27]) using five baseline models. The experimental results demonstrate that models equipped with our strategy significantly outperform their baseline counterparts. For instance, UniMatch integrated with our SESC² achieves 2.03% and 1.26% mIoU improvements over the baseline UniMatch under the 1/16 and 1/8 partition on Pascal VOC 2012, respectively.

Our contributions can be summarized as follows:

- We analyze the limitation of the labeled data in SSSS task, which is difficult to provide rich contextual correlations, thus providing a new perspective for building more effective semantic segmentation models.
- We propose a strategy called Seeking Essential Semantics in Changing Environments, which encourages the model to focus on the intrinsic properties of objects by performing background exchange between labeled and unlabeled data for objects of the same category.
- We validate the effectiveness of our method through extensive experiments on Pascal VOC, Cityscapes, and COCO datasets. Across five strong baseline models, incorporating our approach consistently leads to notable performance improvements, highlighting its effectiveness.

II. RELATED WORK

A. Consistency Regularization via Data Perturbation

Consistency regularization [4], [16] serves as a dominant paradigm in semi-supervised semantic segmentation [17], [28], predicated on the core assumption that model predictions for unlabeled data should remain consistent to various perturbations. Early approaches [29], [30], such as Mean Teacher [30] and FixMatch [29], primarily enforced weak-to-strong consistency by leveraging photometric noise or geometric transformations. However, for the dense prediction task of semantic segmentation, simple image-level perturbations often fail to provide sufficient regularization signals to capture complex

intrinsic information from unlabeled data. In this regard, recent object-centric representation learning methods [21], [22] demonstrate that learning representations around the objects, rather than overall scene statistics, helps enhance modeling capabilities under weakly supervised conditions. For instance, Wolf et al. [21] learned object-centric intrinsic information by maintaining the spatial offset relationships of image patches within the same object. However, this class of methods primarily relies on appearance similarity or spatial distribution priors, which is effective in constrained scenarios with uniform content patterns, but struggles to cope with the issue of contextual diversity in complex, multi-object natural scenes. To enhance the diversity of perturbations and encourage the model to learn more robust contextual representations, recent works [4], [12]–[15] have introduced advanced data mixing strategies. Methods like CutMix [13] and CutMix-CD [15] randomly paste rectangular regions from one image onto another, compelling the model to attend to local discriminative regions rather than relying solely on global context. Further advancing this direction, ClassMix [14] utilizes model predictions to mix images at the object level, aiming to preserve the semantic consistency of object boundaries. Despite their empirical success [4], [12]–[15], these methods typically employ random image mixing, which leads to unreasonable sample combinations—such as placing a domestic cat in a pond—thereby confusing the model’s representation learning. In contrast, the novelty of our method lies in constructing diverse contexts with semantic rationality while elevating simple data augmentation strategies to guide the model in overcoming reliance on specific contexts, thereby facilitating the capture of object features that are more robust to context changes.

B. Learning Robust Features beyond Labeled Priors

Given that aggressive data augmentation may induce confusing supervisory signals, recent works [7], [16]–[19], [31] have shifted focus towards the quality control of pseudo-labels, utilizing feature-level attributes to filter out unreliable predictions or rectify supervisory signals. For example, CorrMatch [16]

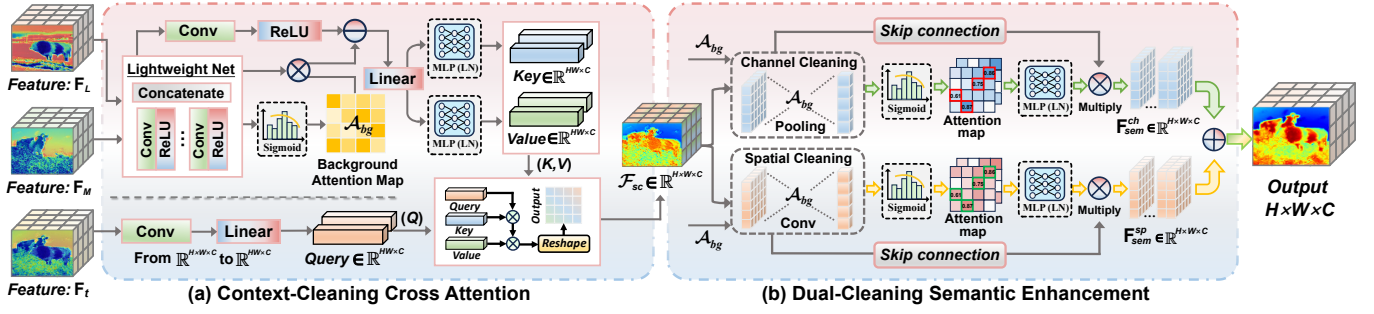


Fig. 4. The architecture of the Semantic Invariant Recombination module. The left part shows the cross-context attention and its core feature shuffle. The right part presents a semantic enhancement module that can take into account both local and global feature enhancement.

employs correlation maps to propagate labels from reliable labeled data to unlabeled queries, while RankMatch [17] evaluates the quality of pseudo-labels based on the rank stability of feature responses under perturbations. Similarly, uncertainty-aware schemes [31] attempt to dynamically adjust supervision weights based on prediction variance. Another line of works [18], [20], [32], [33] emphasize enhancing pseudo-label precision through feature distribution alignment. For instance, methods such as SWSEG [7] and MPMC [18] introduce consistency constraints within student-teacher frameworks to ensure that the feature distribution of unlabeled data aligns with that of the labeled data. DMSI [34] introduces mask-consistent feature alignment to force the model to infer pixel semantics based on the surrounding context, thereby explicitly learning context relationships. Although these methods [7], [16]–[18] are effective, they often fall into a ‘verification logic’ centered on labeled data, leading to excessive reliance on the contexts of labeled data and misinterpreting valid unseen contexts in unlabeled data as statistical anomalies. In contrast, the novelty of our method lies in breaking the constraint of using labeled contexts as the standard for generalization. By constructing various reasonable context transformations, we encourage the model to focus on learning object-related features that are less sensitive to context changes.

III. METHOD

A. Motivation and Overview

Due to limited labeled data and specific context priors [5], models tend to develop a dependency on the contexts of labeled data, leading to recognition failures in the unfamiliar scenes of unlabeled data [4], [16], [28]. To address this, we propose the SESC² framework. By introducing diverse environmental transformations, this framework guides the model to capture foreground attributes that remain relatively stable under changing contexts, thereby achieving robust cross-scene recognition. An overview of our SESC² is illustrated in Fig. 3. It consists of three key components: visual content alteration, semantic invariant recombination, and loss computing, which are detailed in Sec. III-B, III-C, and III-D, respectively.

B. Visual Context Alteration

To decouple objects from context dependencies, we propose Visual Content Alteration (VCA). While prior works [13], [14] have shown that exchanging visual content helps model object-centric representations, their arbitrary pairing often yields

semantically implausible contexts, distracting the model. In contrast, our VCA focuses on exchanging content between foreground entities of the same category, effectively promoting the learning of more generalized object-related semantic features across different backgrounds without introducing counterfactual semantic ambiguities. The key challenge in implementing VCA lies in the limited number of samples of each category within a single batch, which restricts the diversity and randomness of same-category content exchanges. To address the limited diversity of same-category samples within a single batch, we introduce a Class Memory Unit ($\mathcal{M}_N$). $\mathcal{M}_N$ serves as a reservoir, storing up to 50 historical images per category in a First-In-First-Out manner to ensure abundant candidates for the **Writing** and **Reading** processes.

For the Writing process. we first utilize the model [14] to predict the unlabeled images and obtain the segmentation mask (for labeled images, we directly use their corresponding label mask), and then identify the semantic categories contained in each image based on the mask. Then, by computing the number of pixels for each category, we select the category with the largest number of pixels among all semantic categories. The category with the largest number of pixels in the image is considered as the primary category of the image and the image is written into the class memory unit based on this category. The writing process of the $\mathcal{M}_N$ is defined as follows:

$$\mathcal{M}_n^{(t+1)} = \left[(I_L, G_t) \text{ or } (I_U, \tilde{P}_U) \right] + \mathcal{M}_n^{(t)}, \quad (1)$$

where I_L and G_t represent the labeled image and their label to be stored, while I_U and $\tilde{P}_U$ denote the unlabeled image and their mask. n represents the category of the current image. t and $t+1$ refer to the $\mathcal{M}_N$ state from the previous and current write operations. For full-scene image (e.g., the Cityscapes dataset), we randomly select one current class to store into the corresponding class-specific memory slot.

For the Reading process. we first use the model to process the unlabeled images and obtain their mask P_U , and determine the primary semantic class of the image based on the category with the largest number of pixels in the P_U (for labeled images, we obtain the primary semantic class based on the label). Then, we match the obtained category n with the corresponding n -th category in $\mathcal{M}_N$ and select a pair of images and their corresponding labels (or pseudo labels) from the n -th slot for content exchange with the input.

After obtaining the same-category images from $\mathcal{M}_N$, in order to perform content exchange, we first convert the semantic

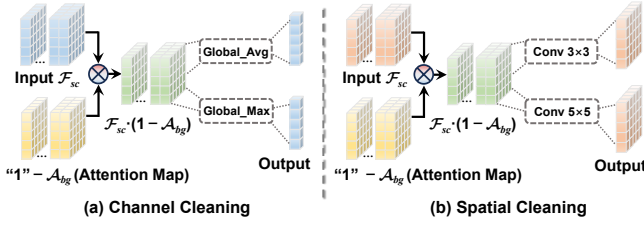


Fig. 5. Flowchart of Channel Cleaning and Spatial Cleaning in Semantic Enhancement. Here, “1” denotes the identity matrix.

map of the input unlabeled image (or the corresponding label for labeled images) into a foreground mask that highlights the key foreground objects as follows:

$$\mathcal{M}_U^f = \begin{cases} 1, & \sum_{h=1}^H \sum_{w=1}^W \text{argmax}(\tilde{P}_U(h, w)) \neq 0 \\ 0, & \sum_{h=1}^H \sum_{w=1}^W \text{argmax}(\tilde{P}_U(h, w)) = 0 \end{cases}, \quad (2)$$

$$\mathcal{M}_L^f = \begin{cases} 1, & \sum_{h=1}^H \sum_{w=1}^W (G_t(h, w)) \neq 0 \\ 0, & \sum_{h=1}^H \sum_{w=1}^W (G_t(h, w)) = 0 \end{cases},$$

where $\mathcal{M}_U^f$ and $\mathcal{M}_L^f$ represent the foreground masks of the label and unlabeled images. Next, we need to perform pixel-wise conversion on G_t and P_U to generate the background mask $\mathcal{M}_L^b$ and $\mathcal{M}_U^b$ that focuses on the background as follows:

$$\mathcal{M}_U^b = \begin{cases} 0, & \sum_{h=1}^H \sum_{w=1}^W \text{argmax}(\tilde{P}_U(h, w)) \neq 0 \\ 1, & \sum_{h=1}^H \sum_{w=1}^W \text{argmax}(\tilde{P}_U(h, w)) = 0 \end{cases}, \quad (3)$$

$$\mathcal{M}_L^b = \begin{cases} 0, & \sum_{h=1}^H \sum_{w=1}^W (G_t(h, w)) \neq 0 \\ 1, & \sum_{h=1}^H \sum_{w=1}^W (G_t(h, w)) = 0 \end{cases}.$$

Finally, based on the foreground and background masks, we perform context exchange between the input and the same-category image I_M provided by the class memory unit:

$$\begin{aligned} I_L^t &= I_L \times \mathcal{M}_L^f + I_M \times \mathcal{M}_L^b, \\ I_U^t &= I_U \times \mathcal{M}_U^f + I_M \times \mathcal{M}_U^b, \end{aligned} \quad (4)$$

where I_L^t and I_U^t are new images generated after content alteration of labeled and unlabeled images. For datasets that do not distinguish between foreground and background, such as the Cityscapes dataset [26], we introduce artificial foreground-background differences by masking out some categories using random masks in Eq. 2 and Eq. 3. Compared with previous works [13], [14], the novelty of Eqs. 2-4 lies in emphasizing the exchange of the most salient same-class content across different images, thereby allowing the same object to appear in diverse yet semantically similar environments. The effectiveness of VCA is validated in Sec. IV-C.

C. Semantic Invariant Recombination

Building on the context alteration described in Sec. III-B, existing works [13], [14] merely treat such transformations as

data perturbation strategies, neglecting to explore how objects maintain semantic invariance amidst changing contexts—a factor critical for recognizing unlabeled data. To address this, as shown in Fig. 4, we propose Semantic Invariant Recombination (SIR), which aims to model semantic representations that are more stable under context changes. Unlike previous method [5] limited to simple feature alignment and enhancement, SIR distinguishes itself by emphasizing the modeling of high-order feature interactions between the original and VCA-transformed views, thereby reconstructing semantic representations that are robust to environmental interference. SIR consists of two components: Context-Cleaning Cross Attention and Semantic Enhancement.

1) *Context-Cleaning Cross Attention*: The Context-Cleaning Cross Attention aims to establish semantic alignment among the transformed view I_L^t , the class memory sample I_M , and the original view I_L (where introducing I_M is crucial as it provides class-consistent supplementary information for I_L^t , which may suffer from semantic ambiguity). While employing cross-attention feature matching to model object representations has proven effective [1], [5], overlooking the background context inherent in the reference features (i.e., F_L and F_M) can interfere with the semantic matching process. Specifically, when the transformed feature F_t serves as a query to match reference features, the contextual noise within these references tends to mislead the attention mechanism into focusing on spurious environmental correlations rather than intrinsic object attributes. Consequently, relying solely on simple cross attention is insufficient to effectively extract object-related semantic representations. To address this, as illustrated in Fig. 4 (a), we propose the Context-Cleaning Cross Attention (CCCA). Unlike traditional methods [1], [5], CCCA incorporates an inverse purification mechanism that adaptively strips away contextual interference from reference features prior to attention computation, thereby compelling the model to perform high-order interactions based on the purified, object-related features. The effectiveness of CCCA is further validated by the ablation study in Sec. IV-C.

First, we use the encoder to encode the images I_L^t (or I_U^t), I_M , and I_L (or I_U) to generate the intermediate features F_t , F_M , and F_L for context-Cleaning cross attention computation: $(F_t, F_M, F_L) = \text{Encoder}(I_L^t, I_M, I_L)$.

After obtaining the intermediate features F_t , F_M , and F_L , we need F_t to search for similar foreground content information in F_M and F_L . Therefore, we linearly encode F_t to generate the *Query* (W^Q) for cross-attention:

$$W^Q = \text{Linear}(\text{Conv}_{1 \times 1}(F_t)), \quad (5)$$

where **Linear** denotes the flattening and linear processing operation. The dimension of W^Q is $HW \times C$.

For F_M and F_L , we need to fuse them to generate the *Value* (W^V) and *Key* (W^K). Specifically, we first concatenate the features F_M and F_L and introduce a lightweight network composed of multiple convolutional layers. This network fuses and projects the reference features into a new feature space, establishing a basis for the subsequent context purification:

$$F_{lw} = \mathcal{N}_{lw}(\text{Concat}(F_M, F_L)), \quad (6)$$

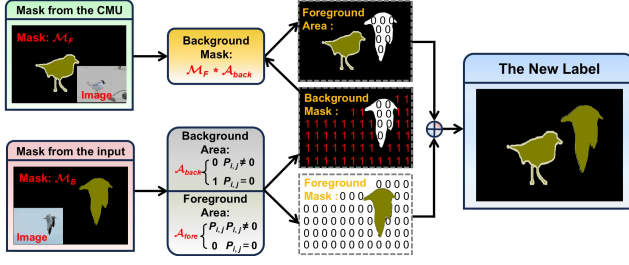


Fig. 6. The generation process of new labels for model training.

where $\mathcal{N}_{lw}(\cdot)$ denotes the lightweight network, and $\mathbf{F}_{lw}$ represents its output feature.

For the feature $\mathbf{F}_{lw}$, we employ a convolutional mapping to generate the feature maps for context cleaning. Simultaneously, we apply a sigmoid function to $\mathbf{F}_{lw}$ to produce the background attention map $\mathcal{A}_{bg}$. $\mathcal{A}_{bg}$ indicates regions within the feature maps that belong to non-essential contextual information. To ensure that effectively captures background features, we leverage the background mask $\mathcal{M}_L^b, \mathcal{M}_U^b$ defined in Eq. 3 to formulate a background constraint loss $\mathcal{L}_{bc}$:

$$\mathcal{L}_{bc} = \mathcal{M}_L^b \log(\mathcal{P}_{bg}) + (1 - \mathcal{M}_L^b) \log(1 - \mathcal{P}_{bg}), \quad (7)$$

where $\mathcal{P}_{bg}$ represents the prediction generated after decoding the background attention map $\mathcal{A}_{bg}$ by the decoder. The analysis of $\mathcal{L}_{bc}$ is provided in the supplementary materials.

Subsequently, we perform the core context cleaning operation. By performing element-wise multiplication between the feature and the background attention map $\mathcal{A}_{bg}$, we estimate the environmental noise component and explicitly subtract it from the input feature:

$$\mathbf{F}_{cc} = \text{Conv}(\text{ReLU}(\mathbf{F}_{lw})) - \mathcal{A}_{bg} \cdot \mathbf{F}_{lw}, \quad (8)$$

where $\mathbf{F}_{cc}$ denotes the feature obtained after context cleaning.

Then, similar to Eq. 5, we flatten $\mathbf{F}_{cc}$ along the spatial dimensions and utilize a Multi-Layer Perceptron (MLP) to project the flattened features, generating the final **Value** ($\mathbf{W}^V$) and **Key** ($\mathbf{W}^K$). After obtaining $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V$, we perform cross-attention computation to model the similar foreground content of $\mathbf{F}_t$ with respect to $\mathbf{F}_M$ and $\mathbf{F}_L$:

$$\mathcal{F}_{sc} \in \mathbb{R}^{H \times W \times C} = \text{Reshape}\left((\mathbf{W}^Q \mathbf{W}^K)^T \mathbf{W}^V\right). \quad (9)$$

Following the standard setting in *ALLSpark* [5], we set the inputs of the Context-Cleaning Cross-Attention to be identical during inference, making it equivalent to self-attention.

2) *Dual-Cleaning Semantic Enhancement*: Following the cross-context attention, it is imperative to further distill discriminative features to bolster the model's representation of discriminative object attributes. In this regard, prevailing paradigms [35] typically employ an attention mechanism. However, these methods often indiscriminately average or convolve the foreground with residual background noise. This counteracts our primary objective of representing the object-centered aspect and poses a risk of re-contaminating the feature representation with residual background interference. To address this, as depicted in Fig.4(a), we propose the Dual-Cleaning Semantic Enhancement (DCSE), whose efficacy is validated in Sec. IV-C. The core innovation of DCSE lies

in the strategic reuse of the background attention map $\mathcal{A}_{bg}$ generated during the CCCA stage, serving as a semantic filter to guide channels and spatial attention to focus exclusively on foreground entities.

First, as shown in Fig. 5, utilizing the $\mathcal{A}_{bg}$, we introduce an identity matrix to invert it into a foreground confidence mask prior to the attention calculation: $\mathbf{M}_{fg} = 1 - \mathcal{A}_{bg}$, to delineate regions containing foreground information within the image. Leveraging the foreground confidence mask $\mathbf{M}_{fg}$, we further purge background interference from $\mathcal{F}_{sc}$:

$$\mathcal{F}_{fg} = \frac{\sum_{i,j} \mathcal{F}_{sc}(i,j) \cdot \mathbf{M}_{fg}(i,j)}{\sum_{i,j} \mathbf{M}_{fg}(i,j) + \varepsilon} \quad (10)$$

where (i,j) denotes spatial coordinates, and ε serves as a term for numerical stability. This design guarantees that the aggregated feature vector is dominated solely by activations from foreground regions, effectively blocking the propagation of background noise into the attention weights. Subsequently, along the channel dimension, we apply both global max and average pooling to $\mathcal{F}_{fg}$ to capture salient responses within the global receptive field. These are then processed by a sigmoid function to generate the channel attention maps $\mathbf{F}_{sem}^{avg}, \mathbf{F}_{sem}^{max}$:

$$\mathbf{F}_{sem}^{avg}, \mathbf{F}_{sem}^{max} = \partial_{\text{sigmoid}}(\mathbf{G}_{avg}(\mathcal{F}_{fg}), \mathbf{G}_{max}(\mathcal{F}_{fg})). \quad (11)$$

For the channel attention maps $\mathbf{F}_{sem}^{avg}, \mathbf{F}_{sem}^{max}$, we employ an MLP to generate channel recalibration weights $\mathcal{W}_{sem}^{avg}, \mathcal{W}_{sem}^{max}$. These weights are then multiplied back into $\mathcal{F}_{sc}$, thereby explicitly enhancing feature channels that exhibit strong responses to the intrinsic attributes of the object:

$$\mathbf{F}_{sem}^{cha} = \mathcal{W}_{sem}^{avg} \cdot \mathcal{F}_{fg} + \mathcal{W}_{sem}^{max} \cdot \mathcal{F}_{fg}. \quad (12)$$

Similarly, at the spatial attention level, we utilize convolutions with kernel sizes of 3 and 5 to process $\mathcal{F}_{fg}$, respectively. This allows for the capture of salient responses within local receptive fields, which are then passed through a sigmoid function to obtain the spatial attention maps $\mathbf{F}_{sem}^{3 \times 3}, \mathbf{F}_{sem}^{5 \times 5}$:

$$\mathbf{F}_{sem}^{3 \times 3}, \mathbf{F}_{sem}^{5 \times 5} = \partial_{\text{sigmoid}}(\mathbf{C}_{3 \times 3}(\mathcal{F}_{fg}), \mathbf{C}_{5 \times 5}(\mathcal{F}_{fg})). \quad (13)$$

For the spatial attention map, an MLP is utilized to generate spatial recalibration weights $\mathcal{W}_{sem}^{3 \times 3}, \mathcal{W}_{sem}^{5 \times 5}$. These weights are applied to $\mathcal{F}_{fg}$, explicitly enhancing local features that correspond to the attributes of the object and generate $\mathbf{F}_{sem}^{spa}$. Finally, $\mathbf{F}_{sem}^{cha}$ and $\mathbf{F}_{sem}^{spa}$ obtained from the channel and spatial attention are fused to produce the semantic feature $\mathbf{F}_{SIR} = \mathbf{F}_{sem}^{cha} + \mathbf{F}_{sem}^{spa}$ for decoding by the decoder.

D. Loss Computation

We employ the standard cross-entropy loss $\mathcal{L}^{ce}(\cdot)$ as the foundational supervision during model training. The total loss function comprises two components: the consistency regularization loss $\mathcal{L}_{cr}$ and our visual context alteration loss $\mathcal{L}_{vca}$. For the consistency regularization loss $\mathcal{L}_{cr}$:

$$\mathcal{L}_{cr} = \frac{1}{2} \mathcal{L}_L^{ce}(p, g) + \frac{1}{2} \mathcal{L}_U^{ce}(p_s, p_w), \quad (14)$$

where $\mathcal{L}_L^{ce}$ denotes the supervised loss for labeled data. p and g represent the model predictions and ground truth for the labeled data, respectively. $\mathcal{L}_U^{ce}$ refers to the unsupervised loss,

TABLE I

COMPARISON WITH SOTA METHODS ON PASCAL VOC DATASET. THE RESULTS OBTAINED USING OUR METHOD ARE HIGHLIGHTED IN **BOLD**. “—” INDICATES THE NULL VALUE, MEANING THAT THE AUTHOR HAS NEITHER PROVIDED THE CORRESPONDING DATA NOR THE CODE FOR REPRODUCTION.

Method.Venues	ResNet-50					ResNet-101				
	1/16(92)	1/8(183)	1/4(366)	1/2(732)	Full(1464)	1/16(92)	1/8(183)	1/4(366)	1/2(732)	Full(1464)
PC ² Seg [36]-[ICCV’21]	56.91	64.63	67.62	70.91	—	57.12	66.34	69.82	73.23	74.17
PS-MT [37]-[CVPR’22]	63.52	68.05	74.19	76.63	78.73	65.77	69.52	76.55	78.49	80.11
ST++ [38]-[CVPR’22]	64.29	70.05	73.95	76.59	77.24	65.28	71.14	74.61	77.42	79.23
U ² PL [31]-[CVPR’22]	65.01	68.32	72.98	75.18	78.66	67.92	69.25	73.77	76.28	79.56
GTA-Seg [39]-[NeurIPS’22]	62.87	71.43	74.35	75.22	77.76	70.09	73.24	75.58	78.47	79.89
AugSeg [28]-[CVPR’22]	64.22	72.18	76.11	77.42	78.84	71.11	75.64	78.93	80.54	81.21
CCVC [40]-[CVPR’23]	66.85	70.99	74.58	77.86	78.59	70.51	77.18	78.73	79.32	81.51
MLLC [19]-[TMM’24]	—	—	—	—	—	70.63	74.34	77.36	79.31	—
MCCL [41]-[TMM’25]	—	75.22	76.45	78.51	79.62	—	—	—	—	—
MGCT [20]-[TMM’25]	—	—	—	—	—	67.20	72.11	75.29	77.51	79.72
UniMatch-[ICCV’23]	71.90	72.51	76.02	77.38	78.70	75.21	77.17	78.83	79.92	81.24
UniMatch + Ours	74.02 +2.12	74.33 +1.82	77.41 +1.39	78.78 +1.40	79.65 +0.95	77.34 +2.13	78.55 +1.38	79.91 +1.08	80.61 +0.69	81.73 +0.49
MPMC-[ECCV’24]	73.51	74.03	77.39	78.52	79.37	77.30	78.58	79.81	80.80	81.67
MPMC + Ours	75.33 +1.82	75.62 +1.59	78.11 +0.72	79.24 +0.72	79.78 +0.41	78.25 +0.95	79.26 +0.68	80.44 +0.63	81.38 +0.58	82.05 +0.38
CorrMatch-[CVPR’24]	71.64	73.85	76.60	77.93	79.27	76.40	78.48	79.42	80.60	81.78
CorrMatch + Ours	73.83 +2.19	75.20 +1.35	77.55 +0.95	78.63 +0.70	79.58 +0.31	78.19 +1.79	79.35 +0.87	80.23 +0.81	81.10 +0.50	82.21 +0.43
RankMatch-[CVPR’24]	71.60	74.61	76.70	78.79	80.02	75.47	77.60	79.80	80.71	82.17
RankMatch + Ours	73.44 +1.84	75.72 +1.11	77.47 +0.77	79.25 +0.46	80.36 +0.34	77.53 +2.06	78.73 +1.13	80.41 +0.61	81.15 +0.44	82.39 +0.22
SWSEG-[CVPR’25]	72.28	74.79	77.28	78.67	79.56	77.05	78.39	79.67	80.92	81.54
SWSEG + Ours	74.35 +2.07	75.72 +0.93	77.96 +0.68	79.28 +0.61	80.02 +0.46	78.24 +1.19	79.17 +0.78	80.21 +0.54	81.23 +0.31	81.85 +0.31

where p_s and p_w denote the predictions under strong and weak augmentations of the original unlabeled image.

The $\mathcal{L}_U^{ce}$ regularizes prediction of the sample under strong perturbations to be the same as that under weak perturbations, which can be formulated as:

$$\mathcal{L}_U^{ce} = \frac{1}{B} \sum \mathbb{I}(\max(p_w) \geq \tau) \cdot \mathbb{H}(p_w, p_s), \quad (15)$$

where B is the batch size for unlabeled data and τ is a pre-defined confidence threshold to filter pseudo-label noise. Following previous work [4], τ is set to 0.95. $\mathbb{H}$ minimizes the entropy between two probability distributions (p_w and p_s):

$$p_w = \mathcal{M}_{ss}(\mathbf{N}_w(I_U)); p_s = \mathcal{M}_{ss}(\mathbf{N}_s(I_U)), \quad (16)$$

where $\mathbf{N}_w(\cdot)$ and $\mathbf{N}_s(\cdot)$ represent the weak and strong perturbations. $\mathcal{M}_{ss}$ is the semantic segmentation model.

For the visual context alteration loss $\mathcal{L}_{vca}$, its role is to constrain the model to model objects in dynamic backgrounds by focusing more on object-related semantic attributes:

$$\mathcal{L}_{vca} = \mathcal{L}^{ce}(\mathcal{M}_{ss}^{sir}(I_L^t, I_l^t)) + \mathcal{L}^{ce}(\mathcal{M}_{ss}^{sir}(I_U^t, I_u^t)), \quad (17)$$

where $\mathcal{M}_{ss}^{sir}$ represents the semantic segmentation model integrated with the Semantic Invariance Recombination module. I_l^t and I_u^t represent the newly generated labels for the labeled and unlabeled images after visual content alteration. During training, we introduce a hyperparameter α for $\mathcal{L}_{vca}$ to balance the contribution of this loss term. A detailed analysis of α and $\mathcal{L}_{vca}$ is provided in the supplementary material.

The generation process of the new labels I_l^t and I_u^t is shown in Fig. 6. Let $\mathcal{M}_B$ denote the background mask provided by the class memory unit ($\mathcal{M}_N$), and $\mathcal{M}_F$ denote the foreground mask provided by input. We first transform the pixel values of $\mathcal{M}_B$ to generate two binary masks: $\mathcal{A}_{fore}$, which focuses on

the foreground, and $\mathcal{A}_{back}$, which focuses on the background:

$$\mathcal{A}_{back} = \begin{cases} 0, P_{i,j} \neq 0 \\ 1, P_{i,j} = 0 \end{cases}, \mathcal{A}_{fore} = \begin{cases} P_{i,j}, P_{i,j} \neq 0 \\ 0, P_{i,j} = 0 \end{cases}, \quad (18)$$

where $P_{i,j}$ denotes the pixel value at position (i, j) in $\mathcal{M}_B$. Based on $\mathcal{A}_{fore}$ and $\mathcal{A}_{back}$, we generate new labels for the Eq. 17 by calculating: $\mathcal{M}_F \times \mathcal{A}_{back} + \mathcal{M}_B \times \mathcal{A}_{fore}$. Compared with [13], [14], the novelty of new labels lies in their ability to constrain the model to learn features of the same object that remain stable across diverse contexts. We analyze the $\mathcal{L}_{vca}$ incorporating the new label in the supplementary material.

IV. EXPERIMENTS

A. Experimental Setup

Datasets and Metric. We conduct experiments on three representative datasets—Pascal VOC [25], Cityscapes [26], and COCO [27]. The VOC consists of 1,464 training images, 1,449 validation images, and 1,456 test images. Following the settings in [4], we augment the VOC training set with 9,118 annotated images from the SBD [42] to form the VOC augmented set. The Cityscapes contains 2,975 training images, 500 validation images, and 1,524 test images. The COCO includes 118K training images and 5K validation images. To quantitatively evaluate performance, we use the mean intersection over union (mIoU) as the evaluation metric.

Implementation Details. All experiments are conducted using 4 NVIDIA Tesla A40 GPUs (45 GB each). For the PASCAL VOC 2012 dataset, we set the initial learning rate to 0.001 and train the model for 80 epochs. For the Cityscapes dataset, we set the initial learning rate to 0.005 and train the model for 240 epochs. Notably, for Cityscapes, we adopt a commonly used strategy from prior work [5], namely sliding

TABLE II
COMPARISON WITH SOTA METHODS ON CITYSCAPES DATASET. THE RESULTS OBTAINED USING OUR METHOD ARE HIGHLIGHTED IN **BOLD**.

Method.Venues	ResNet-50				ResNet-101			
	1/16(186)	1/8(372)	1/4(744)	1/2(1488)	1/16(186)	1/8(372)	1/4(744)	1/2(1488)
AEL [43]-[NeurIPS'21]	74.01	75.82	76.23	—	75.80	77.92	79.07	80.36
ST++ [38]-[CVPR'22]	61.41	70.12	73.25	74.73	67.64	73.43	74.64	77.78
U ² PL [31]-[CVPR'22]	65.01	68.32	72.98	75.18	74.90	76.48	78.51	79.12
AugSeg [28]-[CVPR'22]	73.71	76.42	78.75	79.30	75.25	77.83	79.52	80.40
CCVC [40]-[CVPR'23]	67.89	73.65	76.59	77.74	74.92	76.40	77.31	78.24
MLLC [19]-[TMM'24]	—	—	—	—	75.43	77.92	79.31	80.22
MCCL [41]-[TMM'25]	75.69	76.83	78.20	79.33	—	—	—	—
UniMatch-[ICCV'23]	75.02	76.81	77.53	78.63	76.60	77.95	79.23	79.54
UniMatch + Ours	76.40 +1.38	78.06 +1.25	78.44 +0.91	79.18 +0.55	78.58 +1.98	79.55 +1.60	79.90 +0.67	80.33 +0.79
MPMC-[ECCV'24]	76.80	78.02	78.61	79.48	78.49	79.20	80.89	81.30
MPMC + Ours	77.83 +1.03	78.75 +0.73	79.10 +0.49	79.85 +0.37	79.52 +1.03	79.87 +0.67	81.34 +0.45	81.57 +0.27
CorrMatch-[CVPR'24]	76.32	77.54	78.40	78.85	77.30	78.49	79.41	80.38
CorrMatch + Ours	77.43 +1.11	78.33 +0.79	78.95 +0.55	79.49 +0.64	78.56 +1.26	79.17 +0.68	80.10 +0.69	80.78 +0.40
RankMatch-[CVPR'24]	75.41	77.69	79.20	79.51	77.11	78.59	80.01	80.70
RankMatch + Ours	77.50 +2.09	78.45 +0.76	79.56 +0.36	79.72 +0.21	78.36 +1.25	79.30 +0.71	80.49 +0.48	81.15 +0.45
SWSEG-[CVPR'25]	75.28	76.90	77.86	79.02	76.80	78.61	79.48	79.95
SWSEG + Ours	76.67 +1.39	78.10 +1.20	78.54 +0.68	79.67 +0.65	78.69 +1.89	79.50 +0.89	80.02 +0.54	80.61 +0.66

window evaluation. For the COCO dataset, we set the initial learning rate to 0.001 and train the model for 10 epochs.

Baselines. To verify the effectiveness of our method, we integrate our method into several baseline models: UniMatch [4], MPMC [18], CorrMatch [16], RankMatch [17], and SWSEG [7]. Specifically, MPMC adopts UniMatch as its base framework. To demonstrate the generalization of our method, we experiment with two different backbones, ResNet-50 and ResNet-101, within each baseline. To guarantee the fairness, we adopt the standard implementations of all baseline models. Our strategy is integrated by generating context-altered samples via the VCA module and processing them through the SIR module alongside the original data stream. The model is then optimized using the sum of the baseline's original consistency loss and our semantic invariance loss, with all other details remaining consistent with the original papers.

B. Comparison with State-of-the-Art Methods

Results on the Pascal. As detailed in Tab. I, our method consistently improves performance across various baselines. With the ResNet-50 backbone, it achieves mIoU gains of 1.82%–2.19%, 0.93%–1.82%, 0.68%–1.39%, and 0.46%–1.40% on the 1/16, 1/8, 1/4, and 1/2 splits, respectively, with similar trends observed on ResNet-101. Notably, even against robust baselines like MPMC, CorrMatch, RankMatch, and SWSEG, our method yields significant boosts in scarce-data regimes, improving mIoU by 0.95%–2.13% (1/16) and 0.68%–1.38% (1/8). These results suggest that in scenarios where limited contextual features dominate the labeled data, our method enables the model to reason more based on object-related features when encountering unfamiliar contexts in unlabeled data, thereby achieving more robust predictions.

Results on the Cityscapes. We select the Cityscapes dataset as a critical benchmark due to its highly structured urban

scenes, where objects exhibit strong correlations with their surrounding environments. Quantitative results in Tab. II demonstrate that our method consistently enhances all baselines, with exceptional efficacy in scarce-label settings. For instance, on the competitive UniMatch baseline, we achieve mIoU gains of up to 1.98% (1/16 split) and 1.60% (1/8 split), demonstrating superior robustness compared to advanced methods like MPMC and SWSEG. This validates that our VCA strategy effectively mitigates context-dependent misclassification by forcing the model to attend to fine-grained, robust object features rather than relying on context correlations.

Results on the Pascal augmented and COCO. To further evaluate the generalizability and scalability of our proposed framework, we conducted extensive experiments on the larger-scale Pascal VOC Augmented and the challenging COCO datasets. The quantitative results are summarized in Tab. III. On the larger-scale VOC Augmented, our method consistently boosts strong baselines, achieving 1.73% and 1.24% mIoU gains for UniMatch and CorrMatch under the 1/16 partition, confirming effective scalability without performance saturation. On the semantically complex COCO dataset, our strategy proves equally robust: applying it to UniMatch yields substantial improvements of 1.17%–1.59% across low-data regimes (1/512 to 1/128), while RankMatch sees a 1.24% leap under the 1/512 split. These results validate that our Class Memory Unit effectively facilitates diverse intra-class content exchange, enabling the construction of more comprehensive and robust feature representations.

Qualitative Evaluation. To visually assess our method's effectiveness, Fig. 7 presents comparisons against state-of-the-art baselines in challenging unseen scenarios, serving as a stress test for reliance on contextual priors. As shown in the first two rows of Fig. 7, standard baselines (e.g., FixMatch) struggle to segment a dog positioned in a desert. This failure stems from strong contextual bias, where models likely learn

TABLE III
COMPARISON WITH SOTA METHODS ON PASCAL VOC *augmented* (321×321) AND COCO DATASETS.

Method.Venues	Pascal <i>augmented</i> dataset				Method.Venues	COCO dataset			
	1/16(662)	1/8(1323)	1/4(2646)	1/2(5291)		1/512(232)	1/256(463)	1/128(925)	1/64(1849)
AEL [43]-[NeurIPS'21]	77.22	77.57	78.06	80.29	PC ² Seg [36]-[ICCV'21]	29.91	37.50	40.12	43.74
PS-MT [44]-[CVPR'22]	75.50	78.20	78.72	79.76	CISC-R [45]-[TPAMI'23]	32.11	40.20	42.33	—
SPEED [46]-[TCSVT'24]	75.28	76.14	77.02	78.25	MKD [47]-[ACM MM'23]	30.24	38.04	42.32	45.50
MGCT [20]-[TMM'25]	75.19	76.71	76.99	—	SPEED [46]-[TCSVT'24]	31.62	37.93	42.21	46.17
UniMatch-[ICCV'23]	76.52	77.01	77.18	77.65	UniMatch-[ICCV'23]	31.86	38.88	44.35	48.17
UniMatch + Ours	78.25 +1.73	78.66 +1.65	78.90 +1.72	79.13 +1.48	UniMatch + Ours	33.45 +1.59	40.18 +1.30	45.52 +1.17	48.93 +0.76
CorrMatch-[CVPR'24]	77.60	77.79	78.31	80.11	CorrMatch-[CVPR'24]	32.60	39.28	44.72	48.73
CorrMatch + Ours	78.84 +1.24	79.12 +1.33	79.27 +0.96	80.36 +0.25	CorrMatch + Ours	33.68 +1.08	40.25 +0.97	45.66 +0.94	48.95 +0.22
RankMatch-[CVPR'24]	77.22	78.34	78.95	80.30	RankMatch-[CVPR'24]	32.46	39.54	44.70	48.55
RankMatch + Ours	78.52 +1.30	78.77 +0.43	79.40 +0.45	80.51 +0.21	RankMatch + Ours	33.70 +1.24	40.28 +0.74	45.73 +1.03	49.11 +0.56



Fig. 7. Visualization comparison of different baselines on the PASCAL VOC 2012 and Cityscapes (1/4 split) before and after applying our method.

spurious correlations linking “dogs” exclusively to indoor or grassy backgrounds, leading to missed detections in unfamiliar textures. In contrast, our method successfully segments the dog, forcing the model to more focus on intrinsic features (e.g., fur and shape) rather than background context. A similar phenomenon is observed in the third and fourth rows of Fig. 7, where a cow appears on a barren dirt road instead of its canonical pastoral setting. Baselines like UniMatch and MPMC fail to recognize the object, likely hallucinating background classes due to the absence of familiar cues. However, equipped with our approach, the models accurately delineate the cow’s boundaries. This demonstrates that our method effectively disentangles semantic objects from context, enabling recognition based on object features rather than contextual shortcuts.

As shown in Fig. 7, in complex scenes such as those in Cityscapes, which involve cluttered backgrounds and multiple co-occurring objects, the baseline methods are generally able to recover the overall structure of major regions, such as roads and vehicles. However, they still tend to produce local confusion in cluttered roadside areas. For example, the grass region on the left side of the figure is consistently misclassified as trees by the baseline models, due to the visual

TABLE IV
ABLATION STUDIES OF KEY MODULES OF SESC² ARE CONDUCTED TO EXAMINE THE CONTRIBUTIONS OF THESE MODULES.

Module					Model	Dataset: Pascal VOC			
VCA	SIR	CCCA	DCSE	—	—	1/16	1/8	1/4	1/2
×					Model-1	60.27	67.48	71.55	73.78
	×				Model-2	60.75	67.68	71.74	73.82
		×			Model-3	61.54	67.70	72.17	74.18
			×		Model-4	61.31	67.22	72.02	74.12
✓	✓	✓	✓		Model-5	62.63	68.74	72.52	74.69

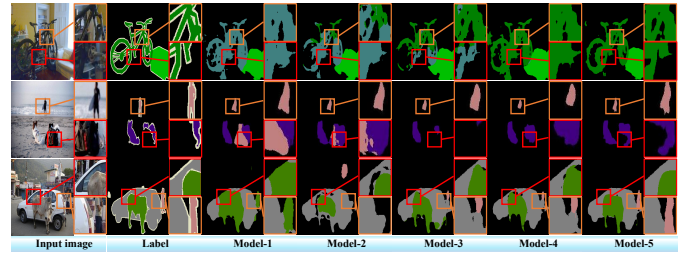


Fig. 8. Visualization results of the ablation studies under the VOC 1/16 split.

similarity between grass and tree regions. In contrast, after incorporating our method, the baseline models exhibit more stable predictions in these complex and ambiguous regions. This improvement can be attributed to the fact that our method constructs diverse background contexts to encourage the model to rely more on object-centric features for recognition, thereby yielding more robust results in complex multi-object scenes.

C. Ablation Studies

We conduct ablation studies to analyze the contributions of key components: Visual Content Alteration (VCA), Semantic Invariance Recombination (SIR) (including Context-Cleaning Cross Attention (CCCA) and Dual-Cleaning Semantic Enhancement (DCSE)). We employ FixMatch [48] (ResNet-50) as the baseline and evaluate five variants: Model-1 (w/o VCA); Model-2 (w/o SIR); Model-3 (replaces CCCA with a standard Concat-Conv operation); Model-4 (w/o DCSE); and Model-5 (Full Method). As detailed in Tab. IV, Model-5 consistently outperforms all variants. The degradation in Model-1 underscores the importance of VCA in shifting focus to intrinsic object features. The drop in Model-2 highlights SIR’s critical role in capturing object-centered attributes. Furthermore, the

TABLE V
STATISTICAL ANALYSIS OF THE PERFORMANCE OF THE PROPOSED METHOD ACROSS DIFFERENT BASELINES (BACKBONE: RESNET-101).

Method.Venues	Pascal VOC dataset			
	1/16(92)	1/8(183)	1/4(366)	1/2(732)
UniMatch	75.05±0.49	77.35±0.41	78.92±0.35	79.68±0.28
UniMatch + Ours	77.12±0.31	78.50±0.24	79.88±0.19	80.71±0.15
CorrMatch	76.12±0.55	78.41±0.46	79.33±0.26	80.44±0.18
CorrMatch + Ours	78.23±0.25	79.30±0.21	80.17±0.20	81.13±0.11
RankMatch	75.79±0.64	77.48±0.47	79.57±0.36	80.62±0.23
RankMatch + Ours	77.48±0.35	78.78±0.26	80.45±0.21	81.18±0.18
Method.Venues	Cityscapes dataset			
	1/16(186)	1/8(372)	1/4(744)	1/2(1488)
UniMatch	76.28±0.57	78.02±0.42	78.95±0.34	79.37±0.25
UniMatch + Ours	78.62±0.28	79.37±0.24	79.94±0.18	80.25±0.17
CorrMatch	77.14±0.45	78.56±0.33	79.51±0.31	80.30±0.27
CorrMatch + Ours	78.37±0.38	79.23±0.20	80.07±0.12	80.84±0.08
RankMatch	76.85±0.37	78.66±0.32	79.89±0.29	80.77±0.24
RankMatch + Ours	78.42±0.31	79.25±0.28	80.57±0.25	81.11±0.20
Method.Venues	Pascal augmented dataset			
	1/16(662)	1/8(1323)	1/4(2646)	1/2(5291)
UniMatch	76.22±0.61	76.82±0.42	77.27±0.21	77.50±0.24
UniMatch + Ours	78.09±0.22	78.72±0.18	78.83±0.10	79.22±0.11
CorrMatch	77.72±0.41	77.86±0.23	78.17±0.24	80.21±0.16
CorrMatch + Ours	78.95±0.18	79.06±0.16	79.33±0.12	80.40±0.08
RankMatch	77.39±0.35	78.20±0.27	78.79±0.20	80.36±0.15
RankMatch + Ours	78.44±0.20	78.87±0.11	79.34±0.08	80.47±0.06

inferior performance of Model-3 and Model-4 validates the effectiveness of CCCA for multi-dimensional fusion and DCSE for balancing local-global receptive fields, respectively. Visual comparisons in Fig. 8 offer intuitive insights. Model-5 yields the most accurate segmentation. For instance, in the orange box (Row 1), Model-5 successfully identifies a bicycle within an atypical living room context. In contrast, Model-1 and Model-2, lacking content alteration or semantic recombination, remain sensitive to unfamiliar contextual cues, resulting in misclassification. Additionally, although Model-3 and Model-4 incorporate VCA, the absence of either effective feature fusion (CCCA) or fine-grained enhancement (DCSE) leads to suboptimal structural details, such as blurred contours and incorrect edge localization. These results collectively validate the effectiveness of our proposed components in enhancing adaptability to unfamiliar contexts.

Supplementary Material. Due to space limitations in the main text, we provide the following details in the Supplementary Material: (1) A comparison between our method and Transformer-based methods. (2) Analysis of whether the proposed method increases training difficulty. (3) A comparison of the Dual-Cleaning Semantic Enhancement with similar methods. (4) Analysis about the Capacity of the Class Memory Unit. (5) Analysis of the proposed method and background attention map via CAM. (6) Feasibility analysis of VCA as a data augmentation strategy. (7) Performance validation of the proposed method on large model-based methods. (8) Analysis of the proposed loss functions and hyperparameter α .

D. Statistical Analysis of Model Performance

Considering that variations in model initialization may introduce randomness to the experimental results, we con-

ducted multiple repeated experiments to enhance the statistical significance and reproducibility of our findings. Specifically, across the 1/16, 1/8, 1/4, and 1/2 data splits of the Pascal VOC, Cityscapes, and Pascal VOC Augmented datasets, we independently trained the baseline models—both with and without our proposed method—three times. The evaluation results are reported in the format of mean $\pm$ standard deviation. The experimental results are detailed in Tab. V. As shown in Tab. V, models equipped with our method consistently outperform their corresponding baselines in terms of mean mIoU across all datasets and split settings. Furthermore, they exhibit smaller standard deviations, even under extremely limited label conditions (e.g., 1/16 and 1/8 splits). These results not only confirm the reproducibility of the performance improvements, but also verify that our method can decouple the model from contextual dependencies on scarce labeled data, thereby enhancing its robustness.

V. CONCLUSION AND LIMITATION

In this work, we address the task of semi-supervised semantic segmentation by proposing a method that transforms background content to encourage the model to focus on object-centered features that remain robust to contextual changes, thereby enhancing generalization on unlabeled data. Our method consistently improves performance across multiple datasets and mainstream baselines. However, as shown in Tab. I and Tab. II, the performance gain diminishes when the proportion of labeled data increases (e.g., 1/4 and 1/2 splits). We attribute this to the fact that with more labeled data, the model can already learn sufficient contextual and categorical information, reducing the relative advantage of our method. In future work, we aim to extend this principle of “object-centered feature learning” to visual foundation models (e.g., SAM or DINO-based frameworks) to further explore how essential semantics can be distilled from massive-scale unlabeled data without heavy reliance on finetuning.

REFERENCES

- [1] H. Zhong, F. Tang, Z. Chen, H. J. Chang, and Y. Gao, “Amdanet: Attention-driven multi-perspective discrepancy alignment for rgb-infrared image fusion and segmentation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10645–10655, October 2025.
- [2] D. Guo, D.-P. Fan, T. Lu, C. Sakaridis, and L. Van Gool, “Vanishing-point-guided video semantic segmentation of driving scenes,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3544–3553, June 2024.
- [3] H. Zhong, F. Tang, H. J. Chang, X. Zhu, and Y. Gao, “Darkseg: Infrared-driven semantic segmentation for garment grasping detection in low-light conditions,” in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 17680–17687, 2025.
- [4] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, “Revisiting weak-to-strong consistency in semi-supervised semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7236–7246, June 2023.
- [5] H. Wang, Q. Zhang, Y. Li, and X. Li, “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 (CVPR)*, pp. 3627–3636, June 2024.
- [6] F. Zhou, Z. Jiang, H. Zhou, and X. Li, “Smc-nca: Semantic-guided multi-level contrast for semi-supervised temporal action segmentation,” *IEEE Transactions on Multimedia*, vol. 26, pp. 11386–11401, 2024.

- [7] C.-Y. Lu, K. Derakhshandeh, and S. Chaterji, "Improving semi-supervised semantic segmentation with sliced-wasserstein feature alignment and uniformity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20233–20243, June 2025.
- [8] H. Liu, Z. Tan, C. Tan, Y. Wei, J. Wang, and Y. Zhao, "Forgery-aware adaptive transformer for generalizable synthetic image detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10770–10780, June 2024.
- [9] Z. Wang, J. Bao, S. Gu, D. Chen, W. Zhou, and H. Li, "Designdiffusion: High-quality text-to-design image generation with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20906–20915, June 2025.
- [10] H. Kim, S. Ahn, and Y.-D. Seo, "Draw your mind: Personalized generation via condition-level modeling in text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17171–17180, October 2025.
- [11] F. Yang, R. Zhen, J. Wang, Y. Zhang, H. Chen, H. Lu, S. Zhao, and G. Ding, "Heic: Mllm-based hierarchical explainable aigc image implausibility evaluator," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3856–3866, June 2025.
- [12] J. Li, Z. Wang, Y. Gao, and X. Hu, "Exploring high-quality target domain information for unsupervised domain adaptive semantic segmentation," in *Proceedings of the 30th ACM International Conference on Multimedia, MM '22*, p. 5237–5245, Association for Computing Machinery, 2022.
- [13] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [14] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "Classmix: Segmentation-based data augmentation for semi-supervised learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1369–1378, January 2021.
- [15] Q. Shu, X. Zhu, L. Wan, S. Zhao, D. Liu, L. Peng, and X. Chen, "Cutmix-cd: Advancing semi-supervised change detection via mixed sample consistency," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–15, 2025.
- [16] B. Sun, Y. Yang, L. Zhang, M.-M. Cheng, and Q. Hou, "Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3097–3107, June 2024.
- [17] H. Mai, R. Sun, T. Zhang, and F. Wu, "Rankmatch: Exploring the better consistency regularization for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3391–3401, June 2024.
- [18] P. Howlader, S. Das, H. Le, and D. Samaras, "Beyond pixels: Semi-supervised semantic segmentation with a multi-scale patch-based multi-label classifier," in *Computer Vision – ECCV 2024*, (Cham), pp. 342–360, Springer Nature Switzerland, 2025.
- [19] H. Xiao, Y. Hong, L. Dong, D. Yan, J. Xiong, J. Zhuang, D. Liang, and C. Peng, "Multi-level label correction by distilling proximate patterns for semi-supervised semantic segmentation," *IEEE Transactions on Multimedia*, vol. 26, pp. 8077–8087, 2024.
- [20] K. Hu, X. Chen, Z. Chen, Y. Zhang, and X. Gao, "Multi-perspective pseudo-label generation and confidence-weighted training for semi-supervised semantic segmentation," *IEEE Transactions on Multimedia*, vol. 27, pp. 300–311, 2025.
- [21] S. Wolf, M. Lalit, K. McDole, and J. Funke, "Unsupervised learning of object-centric embeddings for cell instance segmentation in microscopy images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21263–21272, October 2023.
- [22] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, "Object-centric learning with slot attention," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 11525–11538, Curran Associates, Inc., 2020.
- [23] L. Wu, L. Fang, X. He, M. He, J. Ma, and Z. Zhong, "Querying labeled for unlabeled: Cross-image semantic consistency guided semi-supervised semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8827–8844, 2023.
- [24] M. Jovana, M. Brian, W. Jacob, B. Lars, and B. Charles, "Representation learning via invariant causal mechanisms," in *International Conference on Learning Representations (ICRL)*, Oct 2021.
- [25] M. Everingham, S. M. A. Eslami, V. Gool, Luc, Williams, C. K. I., J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, 2015.
- [26] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014* (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), (Cham), pp. 740–755, Springer International Publishing, 2014.
- [28] Z. Zhao, L. Yang, S. Long, J. Pi, L. Zhou, and J. Wang, "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 (CVPR)*, pp. 11350–11359, June 2023.
- [29] B. Zhang, Y. Wang, W. Hou, H. WU, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 18408–18419, 2021.
- [30] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.
- [31] Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le, "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4248–4257, June 2022.
- [32] X. Wang, H. Bai, L. Yu, Y. Zhao, and J. Xiao, "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 (CVPR)*, pp. 3303–3312, June 2024.
- [33] Q. Dong, J. Li, and S. Deng, "Towards semi-supervised dual-modal semantic segmentation," *IEEE Transactions on Multimedia*, vol. 27, pp. 8212–8224, 2025.
- [34] Q. Ma, Z. Zhang, P. Qiao, Y. Wang, R. Ji, C. Liu, and J. Chen, "Dual-level masked semantic inference for semi-supervised semantic segmentation," *IEEE Transactions on Multimedia*, vol. 27, pp. 4029–4042, 2025.
- [35] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [36] Y. Zhong, B. Yuan, H. Wu, Z. Yuan, J. Peng, and Y.-X. Wang, "Pixel contrastive-consistent semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7273–7282, October 2021.
- [37] Y. Liu, Y. Tian, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro, "Perturbed and strict mean teachers for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4258–4267, June 2022.
- [38] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "St++: Make self-training work better for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4268–4277, June 2022.
- [39] Y. Jin, J. Wang, and D. Lin, "Semi-supervised semantic segmentation via gentle teaching assistant," in *Advances in Neural Information Processing Systems*, vol. 35, 2022.
- [40] Z. Wang, Z. Zhao, X. Xing, D. Xu, X. Kong, and L. Zhou, "Conflict-based cross-view consistency for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19585–19595, June 2023.
- [41] J. Yin, T. Chen, G. Pei, H. Liu, Y. Yao, L. Nie, and X. Hua, "Semi-supervised semantic segmentation with multi-constraint consistency learning," *IEEE Transactions on Multimedia*, vol. 27, pp. 6449–6461, 2025.
- [42] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *2011 International Conference on Computer Vision*, pp. 991–998, 2011.
- [43] H. Hu, F. Wei, H. Hu, Q. Ye, J. Cui, and L. Wang, "Semi-supervised semantic segmentation via adaptive equalization learning," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 22106–22118, Curran Associates, Inc., 2021.
- [44] Y. Liu, Y. Tian, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro, "Perturbed and strict mean teachers for semi-supervised semantic seg-

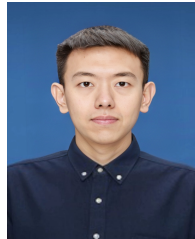
mentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4258–4267, June 2022.

- [45] L. Wu, L. Fang, X. He, M. He, J. Ma, and Z. Zhong, “Querying labeled for unlabeled: Cross-image semantic consistency guided semi-supervised semantic segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8827–8844, 2023.
- [46] X. Lu, L. Jiao, L. Li, F. Liu, X. Liu, and S. Yang, “Self pseudo entropy knowledge distillation for semi-supervised semantic segmentation,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 7359–7372, 2024.
- [47] J. Yuan, J. Ge, Z. Wang, and Y. Liu, “Semi-supervised semantic segmentation with mutual knowledge distillation,” in *Proceedings of the 31st ACM International Conference on Multimedia*, MM ’23, (New York, NY, USA), p. 5436–5444, Association for Computing Machinery, 2023.
- [48] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, “Fixmatch: Simplifying semi-supervised learning with consistency and confidence,” in *Advances in Neural Information Processing Systems*, vol. 33, pp. 596–608, Curran Associates, Inc., 2020.

VI. BIOGRAPHY SECTION



Haifeng Zhong received the B.S. degree in computer science from Shandong University of Traditional Chinese Medicine, Jinan, China, in 2019, and the M.S. degree in computer science from Shandong University of Science and Technology, Qingdao, China, in 2023. He is currently working toward the Ph.D. degree with Jilin University. His research interests include computer vision, image processing and machine learning.



Fan Tang (Member, IEEE) received the B.Sc. degree in computer science from North China Electric Power University, Beijing, China, in 2013, and the Ph.D. degree from Institute of Automation, Chinese Academy of Sciences, Beijing, in 2019. He is an Assistant Professor with the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include computer graphics, computer vision, and machine learning.



Tong-Yee Lee (Senior Member, IEEE) received the Ph.D. degree in computer engineering from Washington State University, Pullman, in May 1995. He is currently a chair professor with the Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan, Taiwan. He leads the Computer Graphics Group, Visual System Laboratory, National Cheng-Kung University (<http://graphics.csie.ncku.edu.tw>). His current research interests include computer graphics, non-photorealistic rendering, medical visualization, virtual reality, and media resizing. He is a senior member of the IEEE Computer Society and a member of the ACM. He also serves on the editorial boards of the IEEE Transactions on Visualization and Computer Graphics, and IEEE Computer Graphics and Applications.



Runyang Feng is a Ph.D. student in the School of Artificial Intelligence at Jilin University. His current research focuses on computer vision, 2D human pose estimation, and video representation learning.



Yixing Gao (Member, IEEE) received the B.Sc. degree in automation from Xi'an Jiaotong University, Xi'an, China, in 2011, and the Ph.D. degree from Imperial College London, London, in 2017. She is an Associate Professor in the School of Artificial Intelligence at Jilin University. Her research interests include humanoid robot, computer vision and human-robot interaction.