Semantic-driven Distillation for Semantic Segmentation with Unknown Classes

Jialei Chen[1], Dongyue Li[1], Chong Yi[1], Xu Zheng[2], Ito Seigo[1], Hiroshi Murase[1], Daisuke Deguchi[1]
[1]Nagoya University, Japan
[2]The Hong Kong University of Science and Technology (Guangzhou)

***Abstract*—Intelligent transportation systems (ITS) rely on semantic segmentation for dense scene understanding and safe decision-making. However, real-world ITS scenarios often involve rare or uncommon objects that may significantly impact decision-making. To handle such cases, models must go beyond closed-set assumptions. Zero-shot semantic segmentation (ZSS) addresses this by allowing models to segment novel classes without labeled examples. To adapt models to these unseen classes, existing approaches typically rely on self-training strategies, where pseudo labels are generated for unlabeled regions based on high-confidence predictions. However, these methods often underutilize the semantic embeddings, which are merely employed to produce pseudo labels, thereby failing to fully exploit CLIP's powerful vision-language alignment capabilities. To address this limitation, we propose Semantic-Driven Distillation (SDD). Specifically, SDD aggregates dense features from a segmentation model into a predicted CLS token via a weighted sum, where the weights are computed based on similarity to the original CLS token from the CLIP visual encoder. It then constructs probability distributions over the predicted and original CLS tokens, as well as the corresponding text embeddings, and aligns these distributions using KL divergence. By leveraging semantic embeddings as a bridge, SDD enables the segmentation model to better align with the CLIP visual encoder, thereby inheriting CLIP's strong vision-language matching capabilities. To further enhance the effectiveness of SDD, we introduce Region-aware Self-Training (RST), which first discovers potential object regions by clustering dense features extracted from CLIP. Within each region, high-confidence predictions are selected as pseudo labels for novel classes. Extensive experiments on standard ZSS benchmarks demonstrate the effectiveness of our proposed approach.**

***Index Terms*—Semantic-driven Distillation, Zero-shot Learning, Semantic Segmentation**

I. Introduction

Intelligent transportation systems (ITS) rely heavily on accurate scene understanding to support applications such as autonomous driving, traffic monitoring, and advanced driver assistance systems (ADAS). Semantic segmentation, which assigns a semantic label to each pixel in an image [1]–[3], plays a crucial role in ITS by providing fine-grained perception of the environment, including roads, vehicles, pedestrians, traffic signs, and lane markings. To enable segmentation models to recognize a wide range of classes, a substantial amount of high-quality annotated data is typically required. Although semantic segmentation models benefit from full supervision, real-world ITS scenarios often involve numerous rare or uncommon object classes that are not present in the training data. These rare classes, such as strollers, traffic cones, or animals crossing the road, may significantly affect driving decisions despite their low frequency. This has led to growing interest in zero-shot semantic segmentation (ZSS), which aims to recognize and segment such unseen classes with only partial supervision. This paradigm substantially reduces the dependence on exhaustive annotations and has emerged as a promising direction for scalable and efficient semantic understanding.

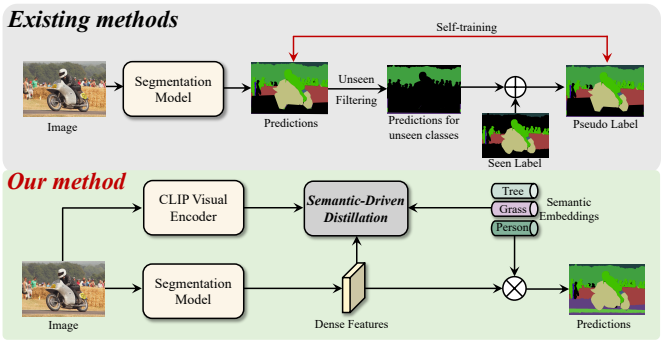

Fig. 1. Comparison between existing methods and ours.

Depending on whether the names of novel classes are available during training, existing zero-shot semantic segmentation methods can be categorized into inductive (class names unknown) and transductive (class names known) settings [4]–[8]. In practice, transductive evaluation is typically performed by first training a segmentation model under the inductive setting for several iterations, followed by applying self-training techniques [4], [9]–[11] to generate pseudo labels for unseen classes and further enhance the segmentation performance. Despite their success, existing methods face two main limitations, including ① an insufficient utilization of the semantic embedding where the semantic embeddings are only used for producing pseudo labels and serving as a classifier; and ② a simple self-training method may overlook the less-dominant classes, leading to suboptimal performance.

To address these two challenges, we introduce Semantic-Driven Distillation (SDD) and Region-aware Self-Training (RST), respectively. By aligning the predicted CLS tokens from dense segmentation features with CLIP's visual and textual embeddings through a semantic bridge formed by text, SDD distills CLIP's vision-language matching capability into the segmentation model, ultimately enabling it to perceive and segment unseen classes. RST enhances the learning of less-dominant unseen classes by discovering region-level semantic structures from CLIP features and selectively gen-

erating pseudo labels within each region. Specifically, SDD first computes the similarity between dense features from the segmentation model and the CLS token of the CLIP visual encoder [12]. This similarity is then used to perform a weighted sum over the dense features, resulting in the predicted CLS tokens. Subsequently, we compute the distributional distance between the semantic embeddings and both the real and predicted CLS tokens, and apply KL divergence to distill knowledge from CLIP into the segmentation model. In addition, to better utilize less-dominant classes, we observe that different classes may occupy different regions within an image. Based on this observation, RST assumes that even though some classes are correctly segmented, they may still exhibit low confidence scores and thus contribute little during training. To ensure these less-dominant yet correctly predicted classes are fully utilized, RST first clusters the dense features extracted from the CLIP visual encoder to identify regions potentially corresponding to different unseen classes. Within each region, we then select the top-K predictions as pseudo labels to guide the training process.

Unlike existing methods that apply self-training only in transductive zero-shot settings [6], [8], [13], resulting in the underutilization of less-dominant classes, the proposed Semantic-Driven Distillation (SDD) leverages all dense features, including those potentially belonging to unseen classes. By using semantic embeddings as a bridge, SDD effectively aligns the segmentation model with the semantic space of CLIP, enabling the close-set segmentation model for zero-shot tasks. Different from conventional knowledge distillation methods [14]–[17], which require the teacher and student to produce the same type of features (*i.e.*, both dense features or both sparse tokens), our method enables cross-type distillation between dense features and sparse tokens. Compared with CLIP-ZSS [7], which directly aligns visual features, our approach leverages text embeddings as an intermediate bridge for knowledge transfer, preserving semantic structure more effectively. Furthermore, unlike traditional self-training methods that select the top $K\%$ highest-confidence predictions as pseudo labels [6], [8], [13], our proposed Region-aware Self-Training (RST) selects pseudo labels within localized regions. This strategy enables better utilization of less-dominant classes and enhances the robustness of the segmentation model in zero-shot scenarios. Our contributions are listed as follows:

– We propose a novel framework consisting of Semantic-Driven Distillation (SDD) and Region-aware Self-Training (RST), which jointly enhance the ability of closed-set segmentation models to generalize to unseen classes.

- By leveraging semantic guidance only during training, our method enables the segmentation model to perform zero-shot semantic segmentation independently, without relying on the CLIP visual encoder during inference.

– Without introducing additional parameters or computational overhead during inference, our method can be flexibly integrated into current powerful segmentation models and achieves state-of-the-art performance on multiple zero-shot segmentation benchmarks.

II. Related Works

A. Zero-shot Semantic Segmentation

Semantic segmentation assigns a label to each pixel in an image, unlike image classification which predicts a single label per image [2], [18]–[22]. Recently, zero-shot semantic segmentation (ZSS) has gained attention for its goal of segmenting unseen classes absent from training data [4]–[6], [8], [23]. To enable such generalization, many methods employ the CLIP visual encoder [12] due to its strong vision-language alignment, typically combined with visual prompts [24] or adapters [25] to adapt CLIP for dense prediction. However, these approaches often underutilize the semantic embeddings, using them mainly as classifier weights or for pseudo labels, without integrating them directly into the learning process. To address this, we propose Semantic-Driven Distillation (SDD), a novel framework that explicitly aligns segmentation outputs with both CLIP visual and textual embeddings during training. This alignment promotes better vision-language consistency and improves generalization to unseen classes.

B. Knowledge Distillation

Knowledge distillation (KD) transfers the capabilities of a stronger teacher model to a weaker student, enabling the latter to achieve competitive performance with reduced complexity [26]. Existing KD methods are typically categorized as logits-based [15], [27], [28], feature-based [29], [30], or relation-based [14], [31]. Logits-based methods align output distributions; feature-based methods match intermediate representations; relation-based methods distill structural relationships, such as similarities among tokens or features. For example, PADing [31] distills similarity relations between CLIP's text embeddings and the student's. Our proposed Semantic-Driven Distillation (SDD) belongs to the logits-based category but supports cross-type distillation between sparse tokens and dense features, unlike most methods requiring matching representation types. In contrast to PromptKD [17], which fine-tunes the text encoder and may weaken CLIP's generality, our method keeps the text encoder frozen, preserving semantic consistency. Compared to Froster [26] and CLIP-ZSS [7], which do not fully exploit semantic structures, SDD introduces textual embeddings as explicit guidance, facilitating more effective vision-language knowledge transfer.

C. Self Training

Collecting large-scale images with high-quality pixel-level annotations is costly and time-consuming. To better utilize limited labeled data, self-training has been widely adopted in semi-supervised learning [32], [33], domain adaptation [34], and weakly supervised learning [35]. Most methods generate pseudo labels by selecting the top $K\%$ most confident predictions [32], [35] or enforcing consistency across models or augmentations [33]. However, in zero-shot settings, these confidence-based strategies tend to overlook less frequent or ambiguous classes, which are often assigned low confidence or appear inconsistently, resulting in biased training. To address this, we propose Region-aware Self-Training (RST), which

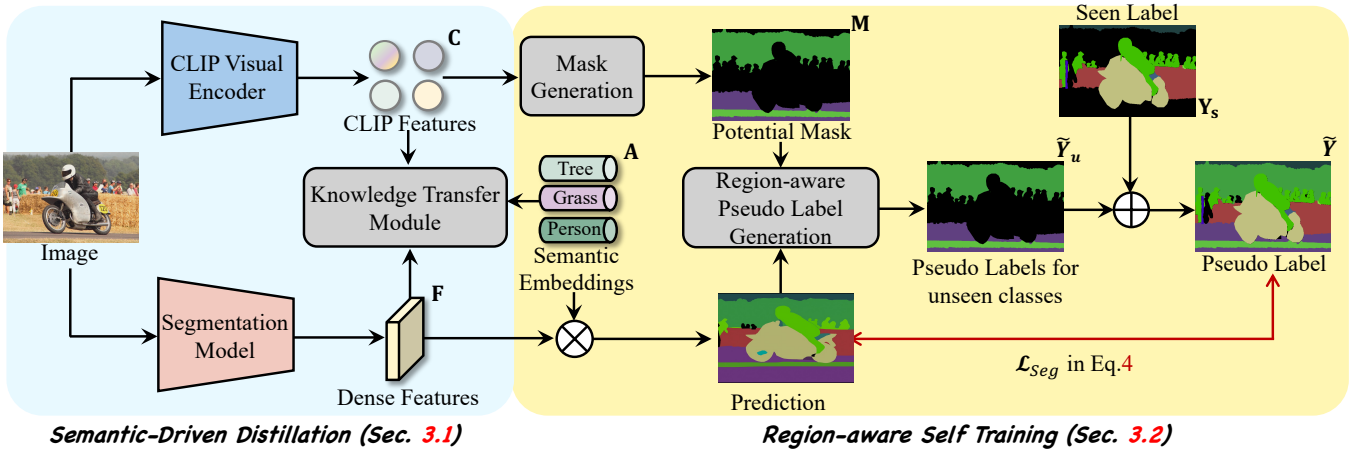

Fig. 2. Overview of the proposed framework. Our method consists of two main components: Semantic-Driven Distillation (SDD) and Region-aware Self-Training (RST). In SDD (left), the segmentation model and the CLIP visual encoder extract dense and sparse features, respectively. The segmentation models are aligned with the CLIP visual encoder via semantic-guided distillation. In RST (right), CLIP features are used to generate region masks, which find potential areas where different unseen classes may appear. Then, the potential masks along with the pixel-level predictions are fed into the proposed region-aware pseudo label generation model to produce the pseudo labels for unseen classes. These pseudo labels are then fused with seen-class ground truth annotations to supervise the model.

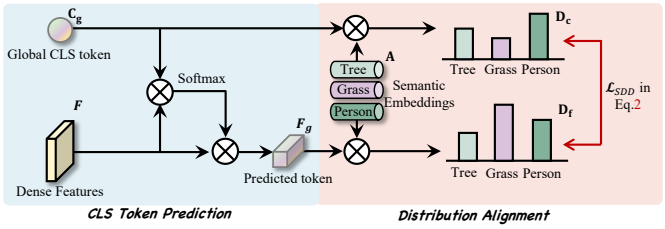

Fig. 3. Overview of knowledge transfer module

identifies spatial regions likely to contain different classes and selects the top-$K\%$ predictions within each region. This region-wise selection increases the inclusion of underrepresented classes, leading to more balanced and effective learning.

III. Method

Preliminary. Before presenting our method, we first define the transductive Zero-shot Semantic Segmentation (ZSS) setting. Formally, let $\mathcal{D} = \{\mathbf{I}^i, \mathbf{Y}_s^i\}_{i=1}^{O}$ denote a training dataset, where $\mathbf{I}^i$ is an input image, and $\mathbf{Y}_s^i$ is the corresponding pixel-level annotation containing only seen classes, and O indicates the size of the dataset. Let $\mathbf{A} \in \mathbb{R}^{N \times D}$ represent a set of text embeddings for all N classes, where D is the embedding dimension. In the transductive setting, seen and unseen classes are allowed to co-occur within the same image. Since filtering out images that contain unseen classes is generally impractical, only the pixel annotations corresponding to unseen classes are all changed into the same 'ignored', while the images themselves are retained during training. During inference, the model is evaluated on both seen and unseen classes simultaneously, and the segmentation output must correctly assign labels across the full set of N classes.

Method Overview. To transfer CLIP's perception of unseen classes into the segmentation model, we propose two complementary components: Semantic-Driven Distillation (SDD) and Region-Aware Self-Training (RST). SDD leverages text embeddings as a semantic bridge to distill CLIP's vision-language matching capability into the segmentation model. It aligns the predicted CLS token, aggregated from dense features, with CLIP's visual and textual representations via distributional matching. RST enhances the utilization of less-dominant unseen classes by clustering CLIP features to identify semantic regions. Pseudo labels are then selectively generated from high-confidence predictions within each region, enabling more balanced and region-aware self-training. The overview of our approach is shown in Fig. 2. First, we feed an input image $\mathbf{X}$ into both the CLIP visual encoder to obtain the CLIP visual features $\mathbf{C}^{(L+1)\times C}$, including L features corresponding to the patches of CLIP and one additional CLS tokens. Meanwhile, we also feed the image into a segmentation model, *e.g.*, segformer [19], to obtain the dense features $\mathbf{F}^{H\times W\times C}$. Together with the semantic embeddings $\mathbf{A}$ extracted from the textual encoder of CLIP, we use the proposed semantic-driven distillation to transfer CLIP's strong vision-language matching capabilities to the segmentation model. To further enhance the utilization of the less-dominant classes, we propose the region-aware self training strategy. Given $\mathbf{C}$, we feed it to the mask generation module proposed by CLIP-ZSS [7] to obtain $\mathbf{M}$, and find potential classes in different areas. We also feed the pixel-level prediction $\mathbf{P}^{N\times H\times W}$ obtained by the inner product between $\mathbf{F}$ and $\mathbf{A}$. Then we feed $\mathbf{P}$ and $\mathbf{M}$ to the region-aware pseudo label generation module to produce the pseudo labels for the unseen classes $\tilde{\mathbf{Y}}_u$ and add the given seen labels together as the final pseudo labels $\tilde{\mathbf{Y}}$. Finally, we use the pseudo labels to supervise the segmentation model. The details of SDD and RST are illustrated in Sec. III-A and the Sec. III-B, respectilvely.

A. Semantic-driven Distillation (SDD)

The core idea of **Semantic-Driven Distillation (SDD)** is to fully leverage CLIP's well-structured semantic embedding

space as an intermediate bridge between the dense visual features extracted by the segmentation model and the CLS token from the CLIP visual encoder. This design enables the transfer of CLIP's powerful vision-language alignment capabilities into the segmentation model, enhancing its semantic consistency and generalization to unseen classes. To enable this idea, we propose a knowledge transfer module as shown in Fig. 3.

Given an input image $\mathbf{X}$, we feed it into both a segmentation model (*e.g.*, SegFormer) and the frozen CLIP visual encoder to obtain two sets of features: (1) the dense feature map $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$ from the segmentation model, and (2) the CLS token representation $\mathbf{C} \in \mathbb{R}^{(1+L)\times C}$ from CLIP, which contains a global CLS token and patch-level representations. We denote the global CLS token as $\mathbf{C}_g \in \mathbb{R}^{1\times C}$ and the dense patch tokens as $\mathbf{C}_d \in \mathbb{R}^{L\times C}$. To establish a semantic connection between the segmentation model and CLIP, we feed $\mathbf{F}$ and $\mathbf{C}$ to the knowledge transfer module. This module first computes the similarity between the segmentation model's dense features $\mathbf{F}$ and the global CLS token $\mathbf{C}_g$. Specifically, we apply a scaled dot-product attention followed by a softmax $\mathbf{W} = \text{Softmax}\left(\frac{\mathbf{C}_g^\top \mathbf{F}}{\sqrt{D}}\right)$ where $\mathbf{W} \in [0,1]^{1\times(H\times W)}$ denotes the normalized attention weights indicating the contribution of each spatial location to the global semantic representation. Using the attention weights $\mathbf{W}$, we aggregate the dense features to obtain a global feature $\mathbf{F}_g \in \mathbb{R}^{1\times C}$: $\mathbf{F}_g = \mathbf{W}\cdot\mathbf{F}^\top$.

To distill CLIP's semantic knowledge into the segmentation model, we apply the distribution alignment. Specifically, we compute the semantic distributions over class prototypes using both the CLIP CLS token and the predicted $\mathbf{F}_g$. Given the semantic embedding matrix $\mathbf{A} \in \mathbb{R}^{N\times C}$ (e.g., from the CLIP text encoder), we calculate the similarity logits and normalize them using a softmax with temperature scaling:

$$\mathbf{D}_c = \text{Softmax}\left(\frac{\mathbf{C}_g\mathbf{A}^\top}{\tau_c}\right), \quad \mathbf{D}_f = \text{Softmax}\left(\frac{\mathbf{F}_g\mathbf{A}^\top}{\tau_f}\right), \tag{1}$$

where τ_c and τ_f are temperature hyperparameters that control the confidence (sharpness) of the resulting distributions. Finally, to enforce that the segmentation model's aggregated features $\mathbf{F}_g$ reflect the same semantic distribution as the CLIP CLS token, we minimize the Kullback-Leibler divergence between the two:

$$\mathcal{L}_{\text{sdd}} = \text{KL}(\mathbf{D}_f \,\|\, \mathbf{D}_c), \tag{2}$$

which guides the segmentation model in capturing the semantic richness encoded in CLIP's global representation and aligning its predictions accordingly.

This semantic-level supervision provides an efficient way to inject vision-language knowledge into dense prediction models, even when textual annotations are sparse or only weakly available. Moreover, the use of semantic embeddings as a bridge allows SDD to naturally handle both seen and unseen classes during zero-shot or open-vocabulary segmentation.

Algorithm 1: Region-aware Self Training (RST)

Input: Dense feature map $\mathbf{C}_d$,
Seen label map $\mathbf{Y}_s$,
Seen class embeddings $\mathbf{A}_s$,
Unseen class embeddings $\mathbf{A}_u$,
Sliding window sizes $\mathcal{K}$,
Top-K% threshold K
Output: Supervision label $\tilde{\mathbf{Y}}$

Step 1: Potential Class Region Mining
Initialize region feature set $\mathcal{C} \leftarrow \emptyset$;
foreach $k \in \mathcal{K}$ **do**
  foreach *position* $(i,j) \in \Omega_k$ **do**
    Extract $k\times k$ region centered at (i,j) from $\mathbf{C}_d$;
    Mask out pixels where $\mathbf{Y}_s[u,v] \in \mathbf{A}_s$;
    Compute average feature over valid pixels, append to $\mathcal{C}$;
Apply K-means clustering on $\mathcal{C}$ and generate region masks $\mathbf{M} \in \mathbb{R}^{U\times H\times W}$;
Step 2: Region-aware Pseudo Label Generation
Compute model feature map $\mathbf{F}$;
Compute unseen class scores: $\mathbf{P}_u = \mathbf{F}\mathbf{A}_u^\top$;
Initialize $\tilde{\mathbf{Y}}_u \leftarrow 0$;
foreach *region* $m \in \boldsymbol{M}$ **do**
  Select top-K% pixels in m based on max score in $\mathbf{P}_u$;
  Assign class label via $\arg\max$ over $\mathbf{P}_u$ to selected pixels ;
Combine $\tilde{\mathbf{Y}}_u$ with seen labels $\mathbf{Y}_s$ to obtain $\tilde{\mathbf{Y}}$;
return $\tilde{\boldsymbol{Y}}$

B. Region-aware Self Training (RST)

Another challenge in conventional zero-shot semantic segmentation is the ineffective utilization of less-dominant classes during self-training, which leads to sub-optimal performance for these classes. To address this issue, we propose **Region-aware Self-Training (RST)**, which includes two key steps: (1) *potential class region mining* and (2) *region-aware pseudo label generation*.

To mine potential class regions, we follow the procedure proposed in CLIP-ZSS [7]. Given the dense visual features $\mathbf{C}_d$ from the CLIP visual encoder, we first initialize cluster seeds by applying multi-scale sliding windows over $\mathbf{C}_d$ to form a region-level feature set $\mathcal{C}_k$ that emphasizes potential unseen classes. Specifically, for each window of size $k\times k$ centered at position (i,j), we compute the average of features excluding those belonging to seen classes $\mathbf{A}_s$, based on the seen ground-truth labels $\mathbf{Y}_s$:

$$\mathcal{C}_k = \left\{ \frac{1}{Z_{i,j}} \sum_{u=i}^{i+k-1} \sum_{v=j}^{j+k-1} \mathbb{1}\big(\mathbf{Y}_s[u,v] \notin \mathbf{A}_s\big)\cdot \mathbf{C}_d[u,v] \,\middle|\, (i,j) \in \Omega_k \right\} \tag{3}$$

$$\Omega_k = \left\{ (i,j) \mid i,j \in \left\{0, \left\lfloor \tfrac{k}{2} \right\rfloor, \ldots, \lfloor H_d - k \rfloor \right\} \right\} \tag{4}$$

TABLE I
COMPARISON WITH PREVIOUS METHODS. **BOLD** DENOTES THE BEST PERFORMANCE, AND UNDERLINE DENOTES THE SECOND BEST PERFORMANCE.

Models	Publish	Backbone	COCO-Stuff			PASCAL Context		
			hIoU	**sIoU**	**uIoU**	**hIoU**	**sIoU**	**uIoU**
SPNet+ST [36]	CVPR19	ResNet101 [37]	30.3	34.6	26.9	-	-	-
ZS5 [4]	NeurIPS19		16.2	34.9	10.6	23.4	27.0	20.7
CaGNet+ST [38]	MM20		19.5	35.6	13.4	-	-	-
STRICT [11]	CVPR21		34.8	35.3	30.3	-	-	-
FreeSeg [39]	CVPR23		45.3	42.2	49.1	-	-	-
MaskCLIP+ [40]	ECCV22		45.0	38.1	54.7	53.3	44.4	66.7
Zzseg [23]	ECCV22	ViT-B [41]	41.5	39.6	43.6	-	-	-
ZegCLIP+ST [6]	CVPR23		48.5	40.7	59.9	54.0	47.2	63.2
CLIP-RC+ST [8]	CVPR24		49.7	42.0	60.8	55.1	48.1	63.2
Ours	-	ViT-B [41]	**51.9**	**44.5**	**62.2**	57.4	52.8	63.0
		Segformer-B4 [19]	51.3	44.1	61.3	**60.3**	**53.2**	**69.7**

where Ω_k denotes the set of valid sliding positions for kernel size $k \in \mathcal{K}$, and $Z_{i,j}$ is the number of valid (unmasked) pixels in the (i, j)-th window. $\mathbf{C}_d$ represents the CLIP dense feature map (excluding the CLS token), and $\mathbb{1}(\cdot)$ is an indicator function used to exclude pixels associated with seen classes. After computing $\mathcal{C}_k$, we apply K-means clustering followed by mask merging [7] to obtain the mask set $\mathbf{M} \in \mathbb{R}^{U\times H\times W}$, where U denotes the number of potential unseen classes. Each binary mask in $\mathbf{M}$ corresponds to a candidate object region in the unannotated areas.

Next, we apply region-aware pseudo label generation. Formally, we compute class scores for unseen classes as $\mathbf{P}_u = \mathbf{F}\mathbf{A}_u^\top$, where $\mathbf{F}$ is the visual feature map from the segmentation model, and $\mathbf{A}_u \in \mathbb{R}^{N_u\times C}$ denotes the semantic embeddings for unseen classes where N_u indicates the number of unseen classes. For each region in $\mathbf{M}$, we select the top-$K\%$ most confident predictions to form the region-aware pseudo labels $\tilde{\mathbf{Y}}_u$. These are then combined with the seen ground-truth labels $\mathbf{Y}_s$ to construct the final supervision mask $\tilde{\mathbf{Y}}$ for training the segmentation model.

Training Loss. The final segmentation loss is computed as:

$$\mathcal{L}(\mathbf{P}, \mathbf{Y}) = \mathcal{L}_{seg}(\mathbf{P}, \tilde{\mathbf{Y}}) + \lambda_s \cdot \mathcal{L}_{sdd}(\mathbf{M}^{\sigma^*}, \mathbf{Y}). \tag{5}$$

where $\mathbf{P} = \mathbf{F}\mathbf{A}^\top$ indicates the pixel-level prediction. $\mathcal{L}_{seg}$ indicates the linear combination of cross-entropy and focal loss [42].

Inference. Since the vision-language matching capability has already been transferred from CLIP to the backbone during training, the model no longer requires CLIP visual encoder at inference time. The backbone has effectively learned to align dense visual features with text embeddings, enabling efficient zero-shot inference without additional computational overhead.

IV. EXPERIMENTS

A. Experimental Setup

Datasets. To evaluate the effectiveness of our method, we conduct zero-shot semantic segmentation (ZSS) experiments on two representative benchmarks: **COCO-Stuff** [43] and **PASCAL Context** [44]. The seen/unseen category splits follow the standard protocol used in previous works [6]. *COCO-Stuff* contains a total of 171 semantic classes, which are divided into 156 seen and 15 unseen classes. The training set consists of 118,287 images, while the test set contains 5,000 images. *PASCAL Context* includes 4,996 training images and 5,104 testing images. For ZSS, 59 classes are selected, with 49 used as seen classes and 10 as unseen.

TABLE II
ABLATION ON PROPOSED METHODS WHERE **B** INDICATES THE BASELINE METHODS (SEGFORMER-B4).

Methods	hIoU	sIoU	uIoU
B	58.0	52.6	64.5
B + SDD	59.3	53.4	66.9
B + SDD + RST	**60.3**	**53.2**	**69.7**

TABLE III
COMPARISON ON ST VS. RST WHERE **B** INDICATES THE BASELINE METHODS (SEGFORMER-B4)

Methods	hIoU	sIoU	uIoU
B + ST	58.2	53.1	64.5
B + RST	**60.4**	**53.4**	**69.5**

Implementation Details. The proposed methods are implemented on the MMsegmentation. The CLIP model applied in our method is based on the ViT-B/16 model and the channel of the output text features is 512. All the experiments are conducted on 4 Nvidia A6000 GPUs and the batch size is set to 16 for all three datasets. For both datasets, the size of the input images is set as 512 $\times$ 512. The iterations are set to 40K and 80K for PASCAL Context and COCO-Stuff. The optimizer is set to AdamW with the default training schedule. For all the datasets, in the first half of the training, we do not apply the RST, and in the rest of the iterations, we apply RST. To evaluate the performance of both seen and unseen classes, we apply the harmonic mean IoU (hIoU) following previous works [6]. The relationship between mIoU and hIoU is $hIoU = \frac{2\cdot sIoU\cdot uIoU}{sIoU+uIoU}$ where $sIoU$ and $uIoU$ indicate the mIoU of the seen classes and unseen classes, respectively.

Fig. 4. Visualization comparison with ZegCLIP [6] and our method.

Besides the hIoU, $sIoU$ and $uIoU$ are also applied. For RST, K is set as 75.

B. Experiment Results

Comparison with State-of-the-Art Methods. Table I compares our method with existing state-of-the-art zero-shot semantic segmentation (ZSS). Our method achieves **the best hIoU and uIoU on COCO-Stuff**, outperforming all previous methods. Specifically, with a ViT-B backbone, we obtain 51.9 hIoU and 62.2 uIoU, surpassing the previous best (CLIP-RC+ST) by **+2.0** and **+1.4** respectively. This demonstrates the superior generalization ability of our method to unseen classes. Even when using a different backbone (Segformer-B4), our method still achieves competitive performance (51.3 hIoU, 61.3 uIoU), confirming its robustness across architectures. On PASCAL Context, our method again sets a new state-of-the-art. With the Segformer-B4 backbone, it achieves **60.3 hIoU**, **53.2 sIoU**, and **69.7 uIoU**, exceeding previous top-performing methods (*e.g.*, MaskCLIP+ [40], ZegCLIP+ST [6], CLIP-RC+ST [8]). Notably, we observe significant improvements in uIoU over prior works, indicating that our approach more effectively utilizes unseen class semantics.

Ablation Studies. In the ablation studies, we choose segformer-b4 [19] as the segmentation model and report the performance on Pascal Context [44]. Ablation Study. As shown in Table II, we conduct an ablation study to evaluate the effectiveness of the proposed Semantic-Driven Distillation (SDD) and Region-aware Self-Training (RST). Starting from the baseline model (B), which does not apply any self-training, the introduction of SDD alone brings consistent improvement

TABLE IV
ABLATION STUDIES OF K IN RST

K in RST	hIoU	sIoU	uIoU
50	57.6	53.0	63.2
75	**60.4**	**53.4**	**69.5**
90	41.2	53.3	33.6

TABLE V
ABLATION STUDIES OF τ_f AND τ_c IN SDD

τ_f	τ_c	hIoU	sIoU	uIoU
0.01	0.01	58.0	52.6	64.5
0.05	0.05	59.3	53.4	66.9
0.07	0.07	52.6	53.2	52.0
0.07	0.05	54.9	53.1	56.9
0.05	0.05	47.7	53.1	43.3
0.05	0.01	60.0	52.6	**69.7**
0.07	0.01	**60.3**	**53.2**	**69.7**

across all metrics, boosting hIoU from 58.0 to 59.3 and uIoU from 64.5 to 66.9. When further integrating RST, the performance improves significantly, achieving 60.3 hIoU, 53.2 sIoU, and 69.7 uIoU, which confirms the complementary benefits of SDD and RST.

Comparison with Standard Self-Training. In Table III, we compare our RST with conventional self-training (ST). While ST already provides a slight gain over the baseline, our RST achieves superior results on all metrics, particularly in unseen category performance, increasing uIoU from 64.5 to 69.5. This highlights the effectiveness of region-level pseudo label generation in discovering less dominant classes.

Effect of the Hyperparameter K in RST. We further analyze the impact of the selection ratio K used in RST (Table IV). Setting $K = 75$ yields the best performance, achieving the highest hIoU (60.4) and uIoU (69.5). A smaller value (e.g., $K = 50$) under-selects high-quality predictions, while a larger value (e.g., $K = 90$) introduces noisy labels, which significantly degrades performance. This indicates that a proper balance between confidence and coverage is critical for effective pseudo-labeling.

Effect of the Hyperparameter τ_f and τ_c in SDD. Table V presents an ablation study on the hyperparameters τ_f and τ_c in the SDD module. These parameters control thresholds for feature and class-level filtering, respectively. The results show that setting τ_f = 0.07 and τ_c = 0.01 achieves the best overall performance, with the highest hIoU (60.3) and uIoU (69.7). This suggests that a stricter feature-level threshold combined with a more permissive class threshold leads to better generalization to unseen classes. On the other hand, setting both thresholds equally high or low degrades performance, indicating the importance of balancing selectivity between feature and class levels.

Qualitative Analysis. Figure 4 shows qualitative comparisons with ZegCLIP on unseen classes. Our method produces more coherent and accurate segmentation, correctly identifying fine-grained regions such as the elephant, person, and orange. In contrast, ZegCLIP yields fragmented or misclassified regions, highlighting the advantage of our semantic-guided framework.

V. CONCLUSION

Intelligent Transportation Systems (ITS) require robust semantic segmentation to accurately recognize both common and rare objects for safe decision-making. However, conventional models often struggle with unseen classes that were not annotated during training, limiting their applicability in dynamic real-world scenarios. To address this challenge, we presented Semantic-Driven Distillation (SDD) and Region-aware Self-Training (RST), aiming to enhance the zero-shot semantic segmentation (ZSS) capability by fully leveraging the rich semantic information encoded in CLIP. Our approach aligns dense feature distributions and refines pseudo labels at the region level, significantly improving the model's ability to generalize to unseen categories without relying on labeled examples. Extensive experiments on standard ZSS benchmarks demonstrate that our method consistently outperforms existing approaches. We believe that our work not only advances open-world segmentation research but also contributes to building reliable and generalizable perception systems for practical ITS and other safety-critical applications.

ACKNOWLEDGMENT

Support for this work was given by the Toyota Motor Corporation (TMC) and JSPS KAKENHI Grant Number 23K28164 and 24H00733, and JST CREST Grant Number JPMJCR22D1. However, note that this article solely reflects the opinions and conclusions of its authors and not TMC or any other Toyota entity. Computations are done on "Flow" at the Information Technology Center, Nagoya University.

REFERENCES

[1] J. Chen, D. Deguchi, C. Zhang, X. Zheng, and H. Murase, "Frozen is better than learning: A new design of prototype-based classifier for semantic segmentation," *Pattern Recognition*, vol. 152, p. 110431, 2024.
[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.
[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *TPAMI*, 2017.
[4] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez, "Zero-shot semantic segmentation," in *NeurIPS*, 2019.
[5] J. Ding, N. Xue, G.-S. Xia, and D. Dai, "Decoupling zero-shot semantic segmentation," in *CVPR*, 2022.
[6] Z. Zhou, Y. Lei, B. Zhang, L. Liu, and Y. Liu, "Zegclip: Towards adapting clip for zero-shot semantic segmentation," in *CVPR*, 2023.
[7] J. Chen, D. Deguchi, C. Zhang, X. Zheng, and H. Murase, "Clip is also a good teacher: A new learning framework for inductive zero-shot semantic segmentation," *arXiv preprint arXiv:2310.02296*, 2023.
[8] Y. Zhang, M.-H. Guo, M. Wang, and S.-M. Hu, "Exploring regional clues in clip for zero-shot semantic segmentation," in *CVPR*, 2024.
[9] 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*, pp. 4268–4277, 2022.
[10] X. Zhao, N. C. Mithun, A. Rajvanshi, H.-P. Chiu, and S. Samarasekera, "Unsupervised domain adaptation for semantic segmentation with pseudo label self-refinement," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2399–2409, 2024.
[11] G. Pastore, F. Cermelli, Y. Xian, M. Mancini, Z. Akata, and B. Caputo, "A closer look at self-training for zero-label semantic segmentation," in *CVPR*, 2021.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[13] J. C. Ye, Y. Oh, *et al.*, "Otseg: Multi-prompt sinkhorn attention for zero-shot semantic segmentation," in *ECCV*, 2024.

[14] K. Han, Y. Liu, J. H. Liew, H. Ding, J. Liu, Y. Wang, Y. Tang, Y. Yang, J. Feng, Y. Zhao, *et al.*, "Global knowledge calibration for fast open-vocabulary segmentation," in *ICCV*, 2023.

[15] G. Hinton, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[16] Z. Quan, Q. Chen, M. Zhang, W. Hu, Q. Zhao, J. Hou, Y. Li, and Z. Liu, "Mawkdn: A multimodal fusion wavelet knowledge distillation approach based on cross-view attention for action recognition," *TCSVT*, 2023.

[17] Z. Li, X. Li, X. Fu, X. Zhang, W. Wang, S. Chen, and J. Yang, "Promptkd: Unsupervised prompt distillation for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26617–26626, 2024.

[18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.

[19] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *NeurIPS*, 2021.

[20] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *CVPR*, 2021.

[21] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *CVPR*, 2022.

[22] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *NeurIPS*, 2021.

[23] M. Xu, Z. Zhang, F. Wei, Y. Lin, Y. Cao, H. Hu, and X. Bai, "A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model," in *ECCV*, 2022.

[24] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *ECCV*, 2022.

[25] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *ICML*, 2019.

[26] X. Huang, H. Zhou, K. Yao, and K. Han, "FROSTER: Frozen CLIP is a strong teacher for open-vocabulary action recognition," in *ICLR*, 2024.

[27] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *ICML*, 2021.

[28] 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.

[29] Z. Quan, Q. Chen, M. Zhang, W. Hu, Q. Zhao, J. Hou, Y. Li, and Z. Liu, "Mawkdn: A multimodal fusion wavelet knowledge distillation approach based on cross-view attention for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 5734–5749, 2023.

[30] T. Sun, H. Chen, G. Hu, and C. Zhao, "Explainability-based knowledge distillation," *Pattern Recognition*, vol. 159, p. 111095, 2025.

[31] S. He, H. Ding, and W. Jiang, "Primitive generation and semantic-related alignment for universal zero-shot segmentation," in *CVPR*, 2023.

[32] X. Hu, L. Jiang, and B. Schiele, "Training vision transformers for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4007–4017, 2024.

[33] J. Chen, C. Fu, H. Xie, X. Zheng, R. Geng, and C.-W. Sham, "Uncertainty teacher with dense focal loss for semi-supervised medical image segmentation," *Computers in Biology and Medicine*, vol. 149, p. 106034, 2022.

[34] X. Zhao, N. C. Mithun, A. Rajvanshi, H.-P. Chiu, and S. Samarasekera, "Unsupervised domain adaptation for semantic segmentation with pseudo label self-refinement," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2399–2409, 2024.

[35] H. Zhang, Y. Su, X. Xu, and K. Jia, "Improving the generalization of segmentation foundation model under distribution shift via weakly supervised adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23385–23395, 2024.

[36] Y. Xian, S. Choudhury, Y. He, B. Schiele, and Z. Akata, "Semantic projection network for zero-and few-label semantic segmentation," in *CVPR*, 2019.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[38] Z. Gu, S. Zhou, L. Niu, Z. Zhao, and L. Zhang, "Context-aware feature generation for zero-shot semantic segmentation," in *ACM MM*, 2020.

[39] J. Qin, J. Wu, P. Yan, M. Li, R. Yuxi, X. Xiao, Y. Wang, R. Wang, S. Wen, X. Pan, *et al.*, "Freeseg: Unified, universal and open-vocabulary image segmentation," in *CVPR*, 2023.

[40] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *ECCV*, 2022.

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

[42] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017.

[43] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *CVPR*, 2018.

[44] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *CVPR*, 2014.