CLIP-to-Seg Distillation for Zero-shot Semantic Segmentation

Jialei Chen*, *Student Member, IEEE,* Zhenzhen Quan, Chenkai Zhang, Xu Zheng, *Student Member, IEEE,* Daisuke Deguchi, *Member, IEEE,* Hiroshi Murase, *Life Fellow, IEEE*

***Abstract*—CLIP has greatly advanced zero-shot segmentation by leveraging its strong visual-language association and generalization capability. However, directly adapting CLIP for segmentation often yields suboptimal results due to inconsistencies between image and pixel-level prediction objectives. Additionally, merely combining segmentation and CLIP models often leads to disjoint optimization, introducing significant computational overhead and additional parameters. To address these issues, we propose a novel CLIP-to-Seg Distillation approach, incorporating global and local distillation to flexibly transfer CLIP's powerful zero-shot generalization capability to existing closed-set segmentation models. Global distillation leverages CLS tokens to condense segmentation features and distills high-level concepts to the segmentation model via image-level features. Local distillation adapts CLIP's local semantic transferability to dense prediction tasks using object-level features, aided by pseudo-mask generation for latent class mining. To further generalize the CLIP-distilled segmentation model, we generate latent text embeddings for the mined latent classes by coordinating their text embeddings and dense features. Our method equips existing closed-set segmentation models with strong generalization capabilities for open concepts through effective and flexible CLIP-to-Seg distillation. Without relying on the CLIP model or introducing extra inference overhead, our method seamlessly integrates into existing closed-set segmentation models and enables zero-shot capability, achieving state-of-the-art performance on multiple benchmarks.**

***Index Terms*—CLIP-to-Seg Distillation, Latent Class Mining, Zero-shot Learning, Semantic Segmentation**

I. Introduction

In recent years, semantic segmentation has advanced rapidly, benefiting from deep learning technologies. However, conventional semantic segmentation models are heavily data-dependent [1–3], requiring large volumes of annotated images to achieve satisfactory performance. Collecting these images and annotations is both time-consuming and expensive.

To address this challenge, zero-shot semantic segmentation has been proposed and has gained significant attention [4, 5]. In zero-shot semantic segmentation, models are trained on seen classes and must generalize to unseen classes during inference, relying solely on their text descriptions. To accomplish this, inspired by the works that adopt CLIP [6] to do downstream tasks [7–9], existing methods [4, 5] typically utilize vision-language models with strong zero-shot generalization capabilities, such as CLIP [6], for pixel-level segmentation tasks.

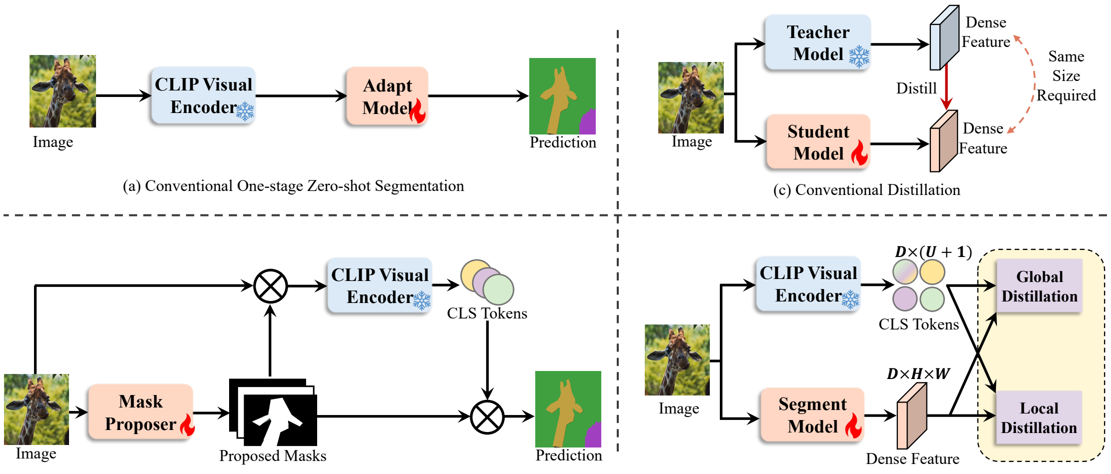
To effectively adapt CLIP for segmentation, existing methods can be categorized into two groups: one-stage methods and two-stage methods, as shown in (a) and (b) of Fig. 1. In one-stage methods [5, 10–12], to maintain CLIP's generalization capability, the adaptation module or trainable prompts are often inserted after the frozen CLIP visual encoder to adapt the dense features for segmentation. Two-stage methods [4, 13] typically require a pre-trained, class-agnostic object proposer to identify latent classes (the classes without labels during training) in an image. These proposals are then fed into the frozen CLIP visual encoder for classification generalization.

Despite their effectiveness, both approaches exhibit inherent limitations. In one-stage methods, CLIP is primarily optimized for capturing global context through the CLS token, but it lacks the spatial information required to capture fine-grained local details necessary for precise segmentation. However, dense prediction tasks prioritize high-quality pixel-level parsing over image-level understanding, creating a mismatch between task requirements and CLIP's capabilities, thus limiting the effectiveness of one-stage methods. Two-stage methods primarily suffer from the disjointed optimization between mask proposal generation and CLIP's class recognition. Additionally, two-stage methods are computationally expensive, as they require both proposal generation and per-proposal classification.

To address the limitations of both approaches, we aim to propose a novel framework that achieves high-quality segmentation without incurring additional computational costs during inference and simultaneously maintains strong zero-shot generalization capabilities. We start by revisiting closed-set segmentation models, which are highly optimized for capturing local details crucial for precise segmentation while achieving high inference speed. However, two key challenges emerge in the context of zero-shot semantic segmentation. First, incomplete annotations prevent the utilization of all the information in an image and tend to bias the seen classes. Second, transferring the vision-language matching capabilities to closed-set segmentation models relies on knowledge distillation techniques. Unfortunately, such approaches typically enforce a consistent representation format, either spatially resolved or non-spatially resolved (see (c) of Fig. 1), which limits the ability to transfer CLIP's knowledge from a single CLS token to dense features within diverse segmentation architectures.

These limitations motivate us to propose CLIP-to-Seg (C2S) distillation which is facilitated by a pseudo mask and latent embedding generation. Different from image classification, semantic segmentation requires both global and local information for segmentation. Therefore, CLIP-to-Seg distillation integrates global and local distillation to transfer CLIP's zero-shot generalization capabilities to the segmentation model as shown in Fig. 1(d). Global distillation adaptively aggregates dense features into one global feature based on their similarity to global CLS tokens which are extracted from the whole image,

Jialei Chen, Chenkai Zhang, Daisuke Deguchi, Hiroshi Murase are with the Graduate School of Informatics, Nagoya University, Nagoya, Japan. Zhenzhen Quan is with the School of Information Science and Engineering, Shandong University, Qingdao, China. Xu Zheng is with AI Thrust, The Hong Kong University of Science and Technology, Guangzhou Campus (HKUST-GZ), Guangzhou, China. * indicates the corresponding author.



(a) Conventional One-stage Zero-shot Segmentation

(c) Conventional Distillation

(b) Conventional Two-stage Zero-shot Segmentation

(d) CLIP-to-Seg Distillation

Fig. 1: Comparisons between CLIP-to-Seg distillation and other methods. **(a)**: Conventional one-stage zero-shot segmentation, **(b)**: Conventional two-stage zero-shot segmentation, where a proposer is trained and frozen CLIP is used to classify the proposals. **(c)**: Conventional knowledge distillation methods require the student and teacher models to be of the same type. **(d)**: Our CLIP-to-Seg distillation transfers the knowledge of CLIP to segmentation models where features with different sizes are aligned and do not rely on CLIP during inference, resulting in high inference performance and efficiency.

and then performs efficient distillation between the global CLS token and the global feature. Local distillation aligns the CLIP CLS tokens extracted from class-specific crops, consisting of both seen and latent classes from the pseudo masks with the dense features from the corresponding regions. Under zero-shot settings, large amounts of areas are unannotated, leading to sub-optimal local distillation. To leverage the information from the entire image, we propose pseudo mask generation. This method utilizes the K-means algorithm to cluster the CLIP dense features from unannotated areas and further refines the results by merging clusters that likely belong to the same class. The merged results are added with the given seen labels to form the pseudo masks. To further increase the capabilities to distinguish between classes, we propose the latent embedding generation to synthesize the text embeddings for the latent classes. By concatenating with the seen text embeddings, these latent embeddings help differentiate features from unannotated areas and annotated areas, enabling further generalization for the closed-set segmentation model.

Unlike existing approaches that adapt the CLIP visual encoder [5, 11] or ensemble with CLIP during inference [4, 10], our method can be seamlessly integrated into existing closed-set segmentation models without relying on the CLIP model or introducing additional computational parameters at inference. Our method also effectively leverages the strengths of powerful task-specific architectures. By decoupling from the fixed CLIP backbone, our approach allows these closed-set segmentation models to be adapted for zero-shot scenarios, thereby significantly enhancing their applicability and performance. Although global-local distillation is a common practice in knowledge distillation [14, 15], our method differs from existing approaches by transferring knowledge from a single token, the CLS token from CLIP, to the dense feature representations of the segmentation model. Our method achieves state-of-the-art performance on multiple zero-shot segmentation benchmarks when incorporated with powerful segmentation models such as Segformer [16] and SegNeXt [17]. In summary, our contributions are:

– We propose a novel CLIP-to-Seg distillation framework that adopts a sparse-to-dense paradigm to transfer CLIP's vision-language matching capabilities to segmentation models.

– We propose a novel pseudo mask generation and latent embedding generation to help the CLIP-distilled segmentation model generalize well on unseen classes.

– Our method introduces no additional parameters or computational overhead, while being fully plug-and-play with existing closed-set segmentation models for zero-shot capabilities, achieving state-of-the-art performance on multiple benchmarks.

II. Related Work

Closed-set Semantic Segmentation: Closed-set segmentation assumes fully annotated images and focuses on the performance of predefined categories within a specific dataset. Existing methods are typically divided into pixel-level classification and mask-level classification. In pixel-level classification, FCN [18], the first fully convolutional network for end-to-end semantic segmentation, established the paradigm. Since FCN, many works, *e.g.*, DeepLab series [19, 20], Deformable convolution [21], aim to enlarge the receptive field to further improve the performance of pixel-level methods. With the introduction of ViT [22], many approaches [16, 17, 23] replaced the conventional convolutional backbone with self-attention-based models, achieving remarkable performance. An alternative approach treats semantic segmentation as a mask classification task. Mask2Former [24] and MaskFormer [25] are notable examples of this approach. Specifically, these models first generate queries corresponding to latent classes. These queries are then decoupled to perform classification and mask prediction tasks separately. Our method is applied to the more challenging task of zero-shot segmentation, which requires fewer annotations than closed-set segmentation.

Knowledge Distillation: Knowledge distillation aims to transfer the capability of a larger teacher model to a student model for comparable performance to the teacher model with a smaller model size [26]. Existing methods are categorized into logits-based [27–29], feature-based [30, 31], and relation-based approaches [14, 32]. With the rapid development of vision-language models [6, 33, 34], certain methods aim to distill vision-language matching capabilities into other models [26, 30, 35]. Although global-local knowledge distillation has been explored in prior works [14, 15], the novelty of our approach lies in distilling knowledge from a single token, namely, the CLS token from CLIP, into dense features of the segmentation model. This contrasts with existing methods that require both teacher and student to share the same feature structure, *i.e.*, either dense-to-dense or sparse-to-sparse.

Zero-shot Semantic Segmentation: Since closed-set segmentation requires pixel-level annotations, research focusing on reducing label dependency has gained significant attention. Before the VLMs, *e.g.*, CLIP, several works tried to bridge the gap between vision and language by projecting the features from vision models to the semantic space [36]. The emergence of large-scale VLMs, such as CLIP [6], has revolutionized zero-shot tasks. Due to their impressive zero-shot ability, researchers aim to transfer this ability to downstream tasks. Leveraging efficient tuning methods [37, 38], existing methods are categorized into one-stage and two-stage approaches. One-stage methods introduce trainable parameters or modules to adapt VLMs for semantic segmentation [4, 5, 8, 9, 39–44]. Two-stage methods train a mask-proposer [24] to propose objects in an image and utilize these objects to finetune the VLMs or directly classify them [13, 41, 45, 46]. Besides zero-shot semantic segmentation, open-vocabulary semantic segmentation also aims to generalize to classes that do not appear during training [47–51]. However, unlike zero-shot methods that are trained on partially labeled data and aim to discover unannotated categories within the same dataset, open-vocabulary methods are trained on fully labeled datasets and focus on transferring to new categories in different datasets.

Different from both types of CLIP-adapting paradigms that rely heavily on CLIP during inference, we propose a CLIP-to-Seg distillation method to transfer the vision-language capability to any pixel-level segmentation model, enabling them to employ zero-shot semantic segmentation without CLIP in inference. Although some methods distill the text relationships to the vision space [14, 26], their methods work under a relaxed condition where all the text embeddings can be accessed. Meanwhile, some object detection methods also try to distill the knowledge from CLIP to detection models [52–55]. However, their methods need to train an additional pseudo mask proposer and provide a detailed description of the input image [52, 53] or need to know all the names of classes [54, 55], which violates the setting of zero-shot learning. Besides, some methods [56] leverage CAM-based techniques to generate pseudo masks. However, our method relies solely on clustering without prior knowledge of the number or identity of the classes.

III. Methods

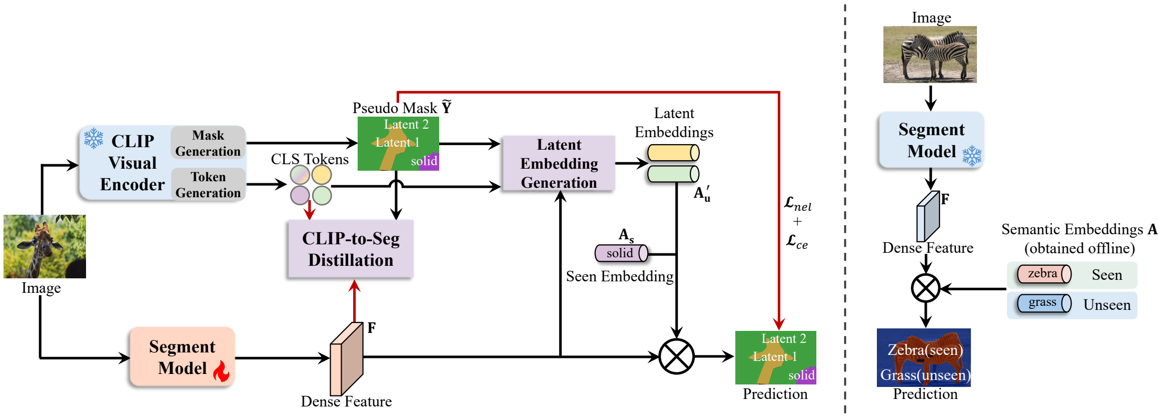
Task Definition: We first define the task of Zero-shot Semantic Segmentation (ZSS). Formally, let $\mathcal{D} = \left\{\mathbf{I}^i, \mathbf{Y}_s^i\right\}_{i=1}^{M}$ represent a dataset, where $\mathbf{I}$ are the input images, $\mathbf{Y}_s$ are the corresponding pixel-level annotations without the annotations of unseen classes, and $\mathbf{A} \in \mathcal{R}^{N \times D}$ is a set of text embeddings for all categories, with N representing the total number of classes and D the dimensionality of the embeddings. The text embeddings $\mathbf{A}$, derived from the CLIP text encoder by applying the prompt template (*e.g.*, "a photo of") with the class name, are partitioned into two disjoint subsets: seen class text embeddings $\mathbf{A}_s \in \mathcal{R}^{N_s \times D}$ and unseen class text embeddings $\mathbf{A}_u \in \mathcal{R}^{N_u \times D}$, where $\mathbf{A}_s \cap \mathbf{A}_u = \varnothing$ and $N_s + N_u = N$. Since seen and unseen classes frequently co-occur in images, removing those containing unseen categories is impractical for training. Therefore, in ZSS, only the annotations for unseen classes are removed. ZSS can be categorized into two settings based on the availability of unseen class text embeddings $\mathbf{A}_u$: *Inductive ZSS*, where unseen class text embeddings are unavailable during training, and *Transductive ZSS*, where unseen class text embeddings are accessible. In both settings, model performance is jointly evaluated on both seen and unseen categories during inference. In this work, we adopt our method for both settings. As some of the annotations are removed and the names or the number of these classes are unknown during training, we define the classes in these areas as latent classes.

A. Basic Idea and Method Overview.

Semantic segmentation requires pixel-wise classification, which differs from image classification which mainly relies on global information. However, existing closed-set segmentation models are limited by their fixed label space, making them difficult to generalize to classes that may not appear in the training dataset. To address this issue, we leverage the strong vision language CLIP and distill its knowledge into segmentation models. Unlike image classification, which only needs global representations, semantic segmentation also demands fine-grained local information. Thus, our knowledge distillation framework consists of both global and local knowledge transfer, and pixel-level supervision to enhance the segmentation:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{I}\in\mathbb{D}}\Big\{\mathcal{L}_g\Big(f_c(\mathbf{I})||f_s(\mathbf{I})\Big)\Big\} \\
&+\mathbb{E}_{\mathbf{I}\in\mathbb{D},\tilde{\mathbf{Y}}_u=M(I)}\Big\{\mathcal{L}_l\Big(f_c(\mathbf{I}|\tilde{\mathbf{Y}}_u)\ ||\ f_s(\mathbf{I}|\tilde{\mathbf{Y}_u})\Big)\Big\} \\
&+\mathbb{E}_{\mathbf{I},\mathbf{Y}_s\in\mathbb{D}}\Big\{\mathcal{L}_l\Big(f_c(\mathbf{I}|\mathbf{Y}_s)\ ||\ f_s(\mathbf{I}|\mathbf{Y}_s)\Big)\Big\} \\
&+\mathbb{E}_{\mathbf{I},\mathbf{Y}_s\in\mathbb{D},\tilde{\mathbf{Y}}_u=M(I),\tilde{\mathbf{Y}}=\mathbf{Y}_s+\tilde{\mathbf{Y}}_u}\Big\{\mathcal{L}_s\Big(f_s(\mathbf{I}),\tilde{\mathbf{Y}}\Big)\Big\}
\end{aligned} \tag{1}$$

where $\mathcal{L}_g$ and $\mathcal{L}_l$ represent the global and local knowledge distillation losses, respectively. $\mathcal{L}_s$ denotes the pixel-level loss that supervises the prediction using pseudo masks $\tilde{\mathbf{Y}}$, which are composed of the provided seen labels $\mathbf{Y}_s$ and the generated pseudo masks $\tilde{\mathbf{Y}}_u$ for unannotated regions. f_c and f_s denote the CLIP model and the segmentation model, respectively. $\mathbf{I}$ denotes the input image, and $\mathbf{Y}_s$ corresponds to the pixel-level annotation mask. M indicates the functions to generate pseudo masks for latent classes in the unannotated areas, and $\tilde{\mathbf{Y}}_u$ indicates the generated pseudo masks. To achieve this,



(a) Training

(b) Inference

Fig. 2: Overview of the CLIP-to-Seg distillation framework. First, the input image is passed through a frozen CLIP visual encoder to obtain both global and local CLS tokens, as well as pseudo masks, which consist of the given seen labels and generated masks for latent classes. The same image is then passed into a trainable segmentation model to extract dense features. CLIP's vision-language matching capabilities are transferred through the proposed CLIP-to-Seg distillation. To provide additional supervision for latent classes, we propose a latent embedding generation method to synthesize text embeddings for latent classes. During inference, our method does not introduce any additional modules or parameters to the segmentation model and relies solely on the segmentation model, resulting in high inference efficiency. All text embeddings are derived from the CLIP text encoder by applying a fixed prompt template (e.g., "a photo of a [class name]") to each category.

we propose CLIP-to-Seg distillation, a simple yet effective approach to generalizing any closed-set segmentation model to classes that do not appear in the training. The core idea can be concluded in Eq. 1. **The first term** distills global knowledge from CLIP by aligning the global CLS token with all dense features from the segmentation model. **The second and third terms** focus on local knowledge distillation. However, under zero-shot settings, many image regions remain unannotated, making direct local supervision sub-optimal. To address this, the second term mines latent classes from the unannotated regions and aligns the corresponding local CLS tokens with dense features extracted from these discovered regions. The third term leverages seen-class annotations by masking the input image accordingly and feeding it into the CLIP visual encoder, extracting local CLS tokens and aligning them with the features of the masked regions. Finally, **the fourth term** provides pixel-wise supervision using both ground-truth masks for seen classes and pseudo masks for unannotated regions. Each pixel is treated as an individual prediction target, enabling the model to learn fine-grained semantic distinctions. This supervision can be implemented with standard pixel-level losses, such as cross-entropy or focal loss, depending on the distribution and confidence of the labels. The overview of our method is shown in Fig. 2, we first generate pseudo masks for latent classes $\tilde{\mathbf{Y}}_u$ in unannotated regions by passing an input image through a frozen CLIP visual encoder and clustering the output features (M), as described in Sec. III-B. We then feed the same image into the CLIP visual encoder and a trainable segmentation model to obtain CLS tokens (including those for latent classes) and dense features, serving as teacher and student features, respectively. Next, we apply the proposed CLIP-to-Seg (C2S) distillation between CLS tokens and dense features to transfer CLIP's knowledge to the segmentation model, as illustrated in Sec. III-C (first, second, and third term in Eq. 1). However, relying solely on C2S distillation may lead to suboptimal performance as unannotated areas can not be fully utilized. To address this, we use a latent embedding generation method (Sec. III-D) to synthesize text embeddings for latent classes. These synthetic text embeddings help distinguish latent from seen classes, providing pixel-level supervision for unannotated regions, aided by the pseudo masks (final term in Eq. 1).

B. Pseudo Mask Generation

In zero-shot settings, the annotations of unseen classes are removed, making the input image not fully utilized. To address this issue, we propose pseudo mask generation (M in Eq. 1) to produce the labels that contain both the given seen labels and the pseudo masks for latent classes. Given an input image, we first feed the image into the frozen CLIP visual encoder to obtain the dense features of CLIP $\mathbf{C}_d$ (the output features excluding the first CLS token). Then, we initialize seeds $\mathbf{C}_{init}$ by applying sliding windows of various sizes to average these dense features:

$$\mathbf{C}_{init} = \left\{ \sum_{u=i}^{i+k-1} \sum_{v=j}^{j+k-1} \frac{\mathbf{C}_d[u,v]}{k^2} \middle| \ \mathbf{C}_d = 0 \ if \ \mathbf{Y}_s[u,v] \in \mathbf{A}_s \right\} \tag{2}$$

where $I = \{0, [k/2], \cdots, [H_d - k]\}$, $J = \{0, [k/2], \cdots, [W_d - k]\}$, $[\cdot]$ denotes the rounding operation where $\mathbf{C}_d$ represents the CLIP visual dense features. $i \in \{0, [k/2], [k], ..., [H_d - k]\}$ and $j \in \{0, [k/2], [k], ..., [W_d - k]\}$ denote the stride of the sliding windows. $k \in \mathbf{K}$ indicates the size of different sliding windows. Here, H_d and W_d represent the size of $\mathbf{C}_d$, and $[\cdot]$ denotes the rounding operation. Based on $\mathbf{C}_{init}$, we apply K-Means clustering to the unannotated regions of $\mathbf{C}_d$ and obtain the clustering results $\mathcal{M} \in \mathbb{R}^{U' \times H \times W}$ and the updated seed features $\mathbf{S}_d \in \mathbb{R}^{U' \times D}$ where U' indicates the number

Algorithm 1 Mask Merging Algorithm

```
Input: Clustered masks $\mathcal{M} \in \mathbb{R}^{U' \times H \times W}$, seed features $\mathbf{S}_d \in \mathbb{R}^{U' \times D}$, similarity threshold $\lambda$
Output: Merged mask $\tilde{\mathbf{Y}}_u$
Initialize similarity matrix $\mathbf{G} \leftarrow \cos(\mathbf{S}_d, \mathbf{S}_d^\top)$
Set diagonal elements: $\mathbf{G}_{i,i} \leftarrow -\infty$, for all $i$
Initialize merged mask $\tilde{\mathbf{Y}}_u \leftarrow \varnothing$
$g_{\text{max}} \leftarrow \max(\mathbf{G})$
while $g_{\text{max}} \geq \lambda$ do
    $i \leftarrow \arg\max(\mathbf{G})$        // Index of the highest similarity
    $\mathcal{I} \leftarrow \{j \mid \mathbf{G}[i,j] > \lambda\}$   // Similar masks to be merged
    $m_{\text{merged}} \leftarrow \sum_{j \in \mathcal{I}} \mathcal{M}[j]$
    $\tilde{\mathbf{Y}}_u \leftarrow \tilde{\mathbf{Y}}_u \cup \{m_{\text{merged}}\}$
    $\mathbf{G}[\mathcal{I}, :] \leftarrow -\infty$; $\mathbf{G}[:, \mathcal{I}] \leftarrow -\infty$
    $g_{\text{max}} \leftarrow \max(\mathbf{G})$
end while
return $\tilde{\mathbf{Y}}_u$
```

of unique masks, and D is the number of channels. Finally, we merge the clustering results as described in Algorithm 1. Formally, the algorithm takes the $\mathbf{S}_d$ and $\mathcal{M}$ as input. First, we compute the similarity matrix $\mathbf{G}$ by computing the cosine similarity among the updated seed features $\mathbf{S}_d$. It iteratively selects the most similar pair of masks, determined by the maximum similarity value g_{max} in $\mathbf{G}$, and adds all masks with similarity greater than λ. The added mask is appended to the result set $\tilde{\mathbf{Y}}_u$. Once a mask is added, its similarity values in $\mathbf{G}$ are set to $-\infty$ to prevent selection next time. This process continues until no similarity value exceeds the threshold λ, and the returned $\tilde{\mathbf{Y}}_u$ will serve as the labels for latent classes. Finally, we add the given seen labels $\mathbf{Y}_s$ and the generated labels for latent classes $\tilde{\mathbf{Y}}_u$ as the pseudo masks $\tilde{\mathbf{Y}}$. Moreover, without relying on annotations, the method effectively discovers latent classes through distinct cluster centers, as illustrated in Fig.6 in Sec.IV-D.

C. CLIP-to-Seg Distillation

The core idea of CLIP-to-Seg (C2S) distillation is to align the CLS tokens that contain the vision-language matching capabilities with the dense features from segmentation models. The CLS tokens include two types: global CLS tokens, which are extracted from the whole image, and local CLS tokens, which are extracted from images masked by the labels. While existing methods typically perform knowledge transfer in a dense-to-dense [29, 30] or sparse-to-sparse [26] fashion, our method uniquely operates in a sparse-to-dense manner, where a single CLS token from the CLIP visual encoder is utilized to transfer semantic knowledge to the dense features of segmentation models. Before introducing the CLIP-to-Seg distillation, we first introduce how the CLS token is extracted as shown in the top of Fig. 3. To obtain the global CLS tokens $\mathbf{C}_g \in \mathbb{R}^{1 \times D}$, we simply input the images into the CLIP visual encoder. To obtain the local CLS tokens, we separate the pseudo mask $\tilde{\mathbf{Y}}$ into non-overlapping class-specific masks which are used to mask the input image $\mathbf{I}$ into class-specific masked images $\mathbf{I}_l^{(U+O) \times H \times W}$,

$$\mathbf{I}_l = \mathbf{I} \odot \mathbb{1}(\tilde{\mathbf{Y}} = l), \quad l \in \tilde{\mathbf{Y}}, \tag{3}$$

where $\odot$ indicates the per-pixel multiplication (image masking). Each class-specific masked image $\mathbf{I}_l$ is then passed through the CLIP visual encoder to extract the corresponding local CLS tokens $\mathbf{C}_l \in \mathbb{R}^{(O+U) \times D}$ where O and U indicate the number of seen and latent classes, respectively.

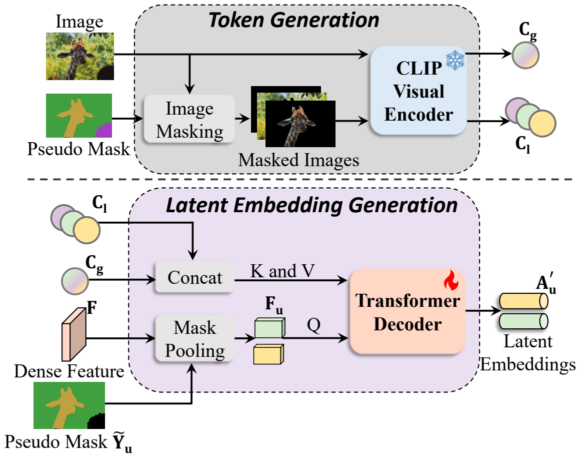


Fig. 3: The overview of token and latent embedding generation.

Once we obtain $\mathbf{C}_g$ and $\mathbf{C}_l$, we can apply the CLIP-to-Seg distillation which consists of two components: global distillation and local distillation. We first introduce global distillation (the first term in Eq. 1), which transfers CLIP's knowledge by aligning global CLS tokens with the global feature. Specifically, as illustrated in the top right of Fig. 4, the input image is passed through a trainable segmentation model to extract dense features $\mathbf{F}^{D \times H \times W}$, where H and W are the height and width of the feature map, respectively. To compute the global feature, $\mathbf{F}$ is reshaped to $D \times L$, where $L = H \times W$. The similarity $\mathbf{W}$ between $\mathbf{F}$ and the global CLS token $\mathbf{C}_g$ is computed as $\mathbf{W} = \text{Softmax}(\frac{\mathbf{C}_g^\top \mathbf{F}}{\sqrt{D}})$, where $\mathbf{W}^{1 \times L} \in [0, 1]$, and the softmax is applied along the second dimension of $\mathbf{W}$. $\mathbf{W}$ represents the similarities between the dense features of the segmentation model and the CLS token, which includes vision-language alignment capabilities. Higher similarity values indicate that the dense features are more semantically aligned with the object described by the CLS token. This similarity is then used to weigh the contributions of each dense feature in generating the global feature $\mathbf{F}_g$, where $\mathbf{F}_g = \mathbf{W} \cdot \mathbf{F}^\top$.

Inspired by the memory buffer mechanism in contrastive learning, which provides additional negative pairs [57], we introduce a CLS token bank to store CLS tokens generated during previous iterations. Specifically, let $\mathcal{V} = \{\mathbf{C}_g^i\}_{i=0}^B$ represent the CLS token bank, where each $\mathbf{C}_g^i$ corresponds to a CLS token collected from earlier training steps and B indicates the size of the bank. In each iteration, before updating the model parameters, we enqueue the current $\mathbf{C}_g$ into $\mathcal{V}$ and dequeue the oldest one. Finally, we align the global feature

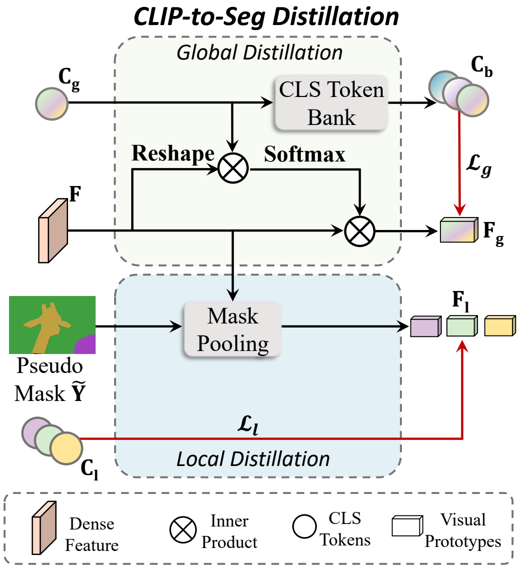


Fig. 4: The process of CLIP-to-Seg distillation.

with the CLS token bank by InfoNCE [1],

$$\mathcal{L}_g = \frac{\exp(\mathbf{F}_g^\top \mathbf{C}_g/\tau)}{\sum_{j=0}^{B} \exp(\mathbf{F}_g^\top \mathbf{C}_j)/\tau)}, \tag{4}$$

where $\mathbf{C}_j \in \mathcal{V}$, and τ denotes the temperature used for contrastive loss. However, due to CLIP's focus on the global context, it may overlook less prominent classes, failing to transfer accurate semantics to the dense features associated with them. To remedy this, we propose the local distillation methods, as shown in the bottom of Fig. 4.

Local distillation (the second and third term in Eq. 1) seeks to transfer semantics overlooked by the global CLS tokens to their corresponding dense features by aligning local features with the local CLS tokens $\mathbf{C}_l$ as shown in the bottom of Fig. 4. Specifically, given the pseudo mask $\tilde{\mathbf{Y}}$, we first mask the dense features from these areas and average the class-specific features to obtain the local features $\mathbf{F}_l \in \mathbb{R}^{(O+U)\times D}$:

$$\mathbf{F}_l = \left\{ \frac{\sum_{H,W} \mathbf{F}[\mathbb{1}(y_i = l)]}{\sum_{H,W}[\mathbb{1}(y_i = l)]} \middle| y_i \in \tilde{\mathbf{Y}} \right\}, \tag{5}$$

where $\mathbb{1}(y_i = l)$ is an indicator function that selects pixels belonging to class l. Finally, given $\mathbf{C}_l$, we apply InfoNCE [1] to align the local features $\mathbf{F}_l$ with the local CLS tokens $\mathbf{C}_l$,

$$\mathcal{L}_l = \sum_{i=0}^{O+U-1} \frac{\exp(\mathbf{f}_i^\top \mathbf{c}_i/\tau)}{\sum_{j=0}^{O+U-1} \exp(\mathbf{f}_i^\top \mathbf{c}_j)/\tau)}, \tag{6}$$

where $\mathbf{f} \in \mathbf{F}_l$ and $\mathbf{c} \in \mathbf{C}_l$. By transferring CLIP's knowledge to segmentation models through C2S distillation, the model's generalization is improved, reducing overfitting to seen classes.

D. Latent Embedding Generation

Although CLIP's vision-language matching capabilities are effectively transferred to segmentation models, the inaccessibility of unseen text embeddings leaves large portions of dense features without pixel-level supervision, resulting in suboptimal optimization of the segmentation model. To address this, we propose latent embedding generation (the fourth term in Eq. 1), which generates synthetic text embeddings for latent classes by calibrating the local features with their corresponding local CLS tokens, as shown in the bottom of Fig. 3.

After obtaining the generated mask for latent classes $\tilde{\mathbf{Y}}_u$, we use Eq. 5 to replace $\tilde{\mathbf{Y}}$ with $\tilde{\mathbf{Y}}_u$ to generate local features $\mathbf{F}_u \in \mathbb{R}^{U\times D}$ for the latent classes. We then feed $\mathbf{F}_u$ into a transformer decoder as query and input the global and local CLS tokens as key and value to generate the latent text embeddings $\mathbf{A}'_u$. The transformer decoder is chosen because the CLS token for latent classes, while possessing vision-language matching capabilities, lacks the discriminative power required for segmentation. Conversely, the local features $\mathbf{F}_u$ for latent classes offer strong discriminative capabilities but lack vision-language matching. The transformer decoder integrates these complementary strengths, producing more representative embeddings for latent classes. The generated text embeddings $\mathbf{A}'_u$ are treated equivalently to seen text embeddings $\mathbf{A}_s$ and are used to serve as the classifier to distinguish between the seen and latent classes. Formally, the class scores for seen and latent categories are $\mathbf{X}_s = \alpha \cdot \mathbf{F}^\top \cdot \mathbf{A}_s$ and $\mathbf{X}_u = \beta \cdot \cos(\mathbf{F}, \mathbf{A}'_u)$, where α and β are hyperparameters that control the scale of latent classes. Note that, since the generated labels $\tilde{\mathbf{Y}}_u$ and generated text embeddings $\mathbf{A}'_u$ for latent classes are not entirely precise, cosine similarity helps prevent overemphasis on misclassification and aids in distinguishing between seen and latent classes. We then concatenate the logits for both seen and unseen classes as $\mathbf{X}_{logits} = \mathrm{cat}(\mathbf{X}_s, \mathbf{X}_u) \in \mathbb{R}^{(N_s+U)\times H\times W}$, where 'cat' denotes concatenation along the class dimension. Finally, $\tilde{\mathbf{Y}}$ is used for pixel-level supervision (the fourth term in Eq. 1) of the dense features by:

$$\mathcal{L}_s = \mathcal{L}_{focal}(\mathbf{X}_{logits}, \tilde{\mathbf{Y}}) + \mathcal{L}_{dice}(\mathbf{X}_{logits}, \tilde{\mathbf{Y}}) + \mathcal{L}_{ce}(\mathbf{X}_{logits}, \tilde{\mathbf{Y}}). \tag{7}$$

where $\mathcal{L}_{focal}$ refers to the focal loss [58], $\mathcal{L}_{dice}$ indicates the DICE loss [5], and $\mathcal{L}_{ce}$ denotes the cross-entropy loss. When only seen classes are present in an image, latent generated text embeddings will not be generated, and only seen text embeddings are used for training.

In this method, semantics are implicitly leveraged through the vision-language alignment previously established by the C2S distillation. While the actual text embeddings for latent classes are inaccessible, the model uses local visual features as a proxy to generate pseudo-text embeddings via a transformer decoder. These visual features, enhanced with local CLS tokens, serve to bridge the semantic gap by approximating the structure of real text embeddings. Thanks to the C2S-induced alignment, even these visually derived embeddings retain semantic structure aligned with CLIP's space, enabling the model to separate seen and latent semantics. Although not equivalent to pure text embeddings, this design allows the model to distinguish latent categories without requiring explicit textual supervision.

E. Training Objective and Inference

Training Objective: To recap, the training objectives of CLIP-to-Seg distillation are:

$$\mathcal{L} = \mathcal{L}_g + \mathcal{L}_l + \mathcal{L}_s, \tag{8}$$

TABLE I: Comparison with state-of-the-art methods under inductive settings where **bold** and underline indicate the best and the second-best performance.

Models	Backbone	PASCAL VOC			COCO-Stuff			PASCAL Context		
		hIoU	**sIoU**	**uIoU**	**hIoU**	**sIoU**	**uIoU**	**hIoU**	**sIoU**	**uIoU**
SPNet [59]	ResNet101 [60]	26.1	78.0	15.6	14.0	35.2	8.7	-	-	-
ZS3 [61]		28.7	77.3	17.7	15.0	34.7	9.5	15.8	20.8	12.7
CaGNet [36]		39.7	78.4	26.6	18.2	33.5	12.2	21.2	24.1	18.5
SIGN [62]		41.7	75.4	28.9	20.9	32.3	15.5	-	-	-
Joint [63]		45.9	77.7	32.5	-	-	-	20.5	33.0	14.9
ZegFormer [4]		73.3	86.4	63.6	34.8	36.6	33.2	-	-	-
Zzseg [13]	ViT-B [22]	77.5	83.5	72.5	37.8	39.3	36.3	-	-	-
ZegCLIP [5]		84.3	91.9	77.8	40.8	40.2	41.4	49.9	46.0	54.6
DeOP [10]		80.8	88.2	74.6	38.2	38.0	38.4	-	-	-
OTSeg+ [64]		87.1	**93.3**	81.6	41.5	41.3	41.8	57.7	**55.2**	60.4
CLIP-RC [11]		88.4	92.8	84.4	41.2	40.9	41.6	51.9	47.5	57.3
Ours	SegNeXt-B [17]	89.3	91.2	87.4	42.5	43.1	41.9	57.6	53.3	62.8
	Setr-B [23]	**90.7**	92.3	**89.2**	**44.8**	**43.8**	**45.9**	56.3	52.4	60.8
	Segformer-B4 [16]	88.7	91.3	86.2	43.9	43.2	44.7	**58.0**	52.6	**64.5**
	ViT-B [22]	90.7	92.1	89.4	43.2	43.6	42.8	57.1	51.9	63.5

TABLE II: Comparison with state-of-the-art methods under tranductive setting where **bold** and underline indicate the best and the second-best performance, and ST indicates self-training [5, 64].

Models	Backbone	PASCAL VOC			COCO-Stuff			PASCAL Context		
		hIoU	**sIoU**	**uIoU**	**hIoU**	**sIoU**	**uIoU**	**hIoU**	**sIoU**	**uIoU**
Zzseg + ST [13]	ViT-B [22]	79.3	79.2	78.1	41.5	39.6	43.6	-	-	-
ZegCLIP + ST [5]		91.1	92.3	89.9	48.5	40.7	59.9	54.0	47.2	63.2
FreeSeg [43]		86.9	82.6	91.8	45.3	42.2	49.1	-	-	-
OTSeg+ ST [64]		**94.4**	**94.3**	**94.3**	49.8	41.4	**62.6**	59.8	**54.0**	67.0
CLIP-RC +ST [11]		93.0	93.9	92.2	49.7	42.0	60.8	55.1	48.1	64.5
Ours + ST	SegNext-B [17]	89.5	91.0	88.0	48.5	43.7	54.5	**60.5**	**54.0**	68.7
	Setr-B [23]	92.4	93.1	91.7	**51.5**	**44.4**	61.3	58.0	51.7	65.9
	Segformer-B4 [16]	89.0	91.6	86.6	**50.8**	43.3	61.4	60.0	52.3	**70.3**
	ViT-B [22]	91.0	92.1	89.9	50.1	43.4	59.2	58.1	52.2	65.6

TABLE III: Efficiency comparisons with other methods.

Method	**Parameter ↓**	**GFLOPS ↓**	**FPS ↑**
Zsseg [13]	61.1 M	1916.7	4.2
ZegFormer [4]	60.3 M	1829.3	6.8
ZegCLIP [5]	**13.8 M**	61.1	25.6
OTSeg+ [64]	**13.8 M**	61.9	22.5
Ours+SegNeXt [17]	32.0 M	**33.5**	**40.9**
Ours+SETR [23]	91.0 M	109.0	20.8
Ours+Segformer [16]	65.7 M	60.7	23.0

TABLE IV: Ablations on proposed modules by Segformer-B4.

Methods	**hIoU**	**sIoU**	**uIoU**
baseline (Segformer-B4)	11.2	41.3	6.4
baseline + latent embedding	20.2	41.4	13.4
baseline + distillation	38.8	41.2	36.6
baseline + distillation + latent embedding	**42.3**	**41.9**	**42.7**

TABLE V: Ablations on global and local distillation.

Distillation		**hIoU**	**sIoU**	**uIoU**
global	local			
-	-	20.2	41.4	13.4
-	✓	36.3	**41.9**	32.1
✓	-	40.8	41.6	40.1
✓	✓	**42.3**	**41.9**	**42.7**

TABLE VI: Ablations on different distillations.

Distillation	**hIoU**	**sIoU**	**uIoU**
Cosine Similarity [65]	37.6	41.4	34.4
L2 Loss [66]	17.8	18.4	17.3
Froster [26]	37.2	41.7	33.6
Our distillation	**42.3**	**41.9**	**42.7**

Inference: Since the vision-language matching capability has already been transferred from CLIP to the backbone during training, we do not rely on CLIP at inference. Instead, we directly use the text embeddings from the text encoder as classifier weights, eliminating the need for latent class embeddings. Moreover, as the number of classes in a dataset is fixed, the text embeddings can be precomputed offline, introducing no additional computational overhead compared to standard segmentation models under the closed-set setting and methods that require additional adapters or visual prompts for adapting CLIP to ZSS, *e.g.*, ZegCLIP [5].

IV. Experiments and Discussions

Dataset: To evaluate the effectiveness of our method, we select three representative benchmarks: PASCAL VOC [67], COCO-Stuff [68], and PASCAL Context [69] to conduct our experiments on zero-shot semantic segmentation (ZSS). The split of seen and unseen categories follows the setting of the previous works [5]. *PASCAL VOC* consists of 10,582 images for training and 1,449 images for validation. Note that we convert the 'background' category to the 'ignored'. For this

TABLE VII: Ablations on global and local CLS tokens in latent embedding generation by Segformer-B4.

Calibration		**hIoU**	**sIoU**	**uIoU**
global	local			
-	-	38.8	41.2	36.6
-	✓	41.9	41.3	42.6
✓	-	41.7	41.4	42.1
✓	✓	**42.3**	**41.9**	**42.7**

TABLE VIII: Ablation on latent embedding and prototype calibrator by Segformer-B4.

Feature	Calibrator	**hIoU**	**sIoU**	**uIoU**
Prototypes	-	41.0	41.5	40.5
Prototypes	MLP	41.5	41.3	41.8
CLS tokens	-	40.2	41.5	38.9
CLS tokens	MLP	40.9	41.5	40.3
Prototypes + CLS tokens	Transformer	**42.3**	**41.9**	**42.7**

dataset, there are 15 seen categories and 5 unseen categories. *COCO-Stuff* contains 171 categories totally. As in previous settings, 171 categories are split into 156 seen and 15 unseen categories. Besides, for the training dataset, there are 118,287 images and 5,000 images for testing. *PASCAL Context* includes 4,996 images for training and 5,104 images for testing. For the zero-shot semantic segmentation task, the dataset is split into 49 seen categories and 10 unseen categories.

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. All the experiments are conducted on 8 V100 GPUs, and the batch size is set to 16 for all three datasets. For all three datasets, the size of the input images is set as 512 × 512. The iterations are set to 20K, 40K, and 80K for PASCAL VOC, PASCAL Context, and COCO-Stuff, respectively. The optimizer is set to AdamW with the default training schedule. In addition, the size of CLS token banks is set as 24. All other settings follow the original segmentation models. To evaluate the performance of both seen and unseen categories, we apply the harmonic mean IoU (hIoU) following previous works [5]. 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 and unseen categories, respectively. Besides the hIoU, $sIoU$ and $uIoU$ are also applied. Frames Per Second (FPS) is tested on RTX 3090.

A. Comparison with State-of-the-arts

We evaluate our method by distilling CLIP into three representative closed-set segmentation models: SegNext, SETR, and Segformer. As shown in Table I, our method consistently outperforms existing state-of-the-art approaches across three challenging benchmarks: PASCAL VOC, COCO-Stuff, and PASCAL Context. For example, our method achieves significant improvements in hIoU over CLIP-RC and OTSeg+, with margins of 2.3%, 3.3%, and 0.3% on the respective datasets. These gains mainly stem from improved generalization to unseen categories. On COCO-Stuff, our method obtains a 4.3% higher uIoU than the best-performing baseline. Similar trends can be observed across all datasets, indicating that our model avoids overfitting to seen categories and better captures transferable semantics. To further validate the robustness of our approach, we also conduct experiments using ViT-B as the backbone. As reported in the last row of Table I, our method still achieves highly competitive performance, confirming its effectiveness regardless of architecture.

Table II further compares our method under the inductive setting with self-training (ST). Our method attains competitive or superior results across all three benchmarks. Notably, we achieve the best hIoU and uIoU on the PASCAL Context dataset, indicating strong generalization under limited supervision. These results collectively demonstrate the advantage of distilling vision-language knowledge into segmentation models and confirm the broad applicability of our framework under both standard and self-training settings.

We also provide a comparison of the computational cost and efficiency of our method with previous methods as shown in Table. III. We use a 512 × 512 image as input, compared with the two-stage methods (first and second row in the table), our method can achieve a much higher inference speed and much lower GFLOPS. Compared with the methods that only add a few trainable parameters, though our trainable parameters are higher than theirs, our method has high flexibility based on the segmentation model. For example, when we choose SegNeXt, an efficient segmentation model, our GFLOPS are nearly 50% of the SOTA one-stage methods, and achieve higher speed.

B. Ablation Studies

To evaluate the effectiveness of our method, we do ablation studies on the COCO-Stuff dataset using 40K training iterations with the same hyperparameters. Segformer-B4 is chosen because of its balance in efficiency and performance. Despite the shorter training schedule, the results also demonstrate the effectiveness of our method.

Ablation studies on the proposed methods: Table IV summarizes the ablation study of our proposed modules using SegFormer-B4. Incorporating the latent embedding module alone yields moderate gains (hIoU from 11.2% to 20.2%, uIoU from 6.4% to 13.4%) by encouraging semantic expansion beyond limited seen categories. However, without explicit vision–language alignment, it struggles to substantially improve generalization to unseen classes. In contrast, our CLIP-to-Seg (C2S) distillation, although implemented with a single loss, is deliberately designed to transfer CLIP's vision–language matching capability from a global token to dense features, enabling pixel-level understanding rather than serving as a simple auxiliary loss. This results in a dramatic boost in zero-shot segmentation performance. When both modules are combined, the model achieves the best overall results. Remarkably, all results are obtained within 40K iterations, half the standard budget, yet surpass prior state-of-the-art methods.

Ablation studies on global and local distillation: We evaluate the individual and combined effects of global and local distillation in Table V. Applying either global or local distillation alone improves performance over the baseline, particularly on unseen classes (uIoU), highlighting their individual effectiveness. Notably, combining both strategies leads to the

TABLE IX: Ablation studies on the size of token banks.

Token bank size	hIoU	sIoU	uIoU
0	41.0	41.5	40.4
2	41.5	41.7	41.4
6	**42.3**	**41.9**	**42.7**
14	40.9	41.3	40.5

TABLE X: Ablation studies on the size of windows in pseudo mask generation.

Window Size	hIoU	sIoU	uIoU
3	41.2	41.3	41.1
7	41.8	**41.3**	42.3
3,7	**42.3**	**41.9**	**42.7**

TABLE XI: Ablation studies on the feature aggregation in global distillation.

Aggregation	hIoU	sIoU	uIoU
mean	38.4	41.1	35.9
max	42.0	41.1	**42.9**
attention	**42.3**	**41.9**	42.7

TABLE XII: Ablation studies on other pseudo mask generation.

Methods	GFLOPs	hIoU	sIoU	uIoU
Mask Proposal [13]	17.6	41.3	41.4	41.2
Panoptic cut [48]	16.5	41.6	41.5	41.7
Ours	17.7	**42.3**	**41.9**	**42.7**

best overall performance across all metrics, demonstrating their complementary benefits in enhancing vision-language alignment. In contrast, approaches lacking such alignment struggle to generalize effectively to unseen categories.

Ablation studies on different distillations: We use contrastive learning to distill the knowledge from CLIP in C2S distillation, here, we try to use different distillation methods to prove the effectiveness of our method as shown in Table VI. First, we change the contrastive distillation to the cosine similarity and find that though the sIoU achieves similar performance, the uIoU drops to 34.4%. Then we change the cosine similarity to the direct L2 loss between the CLS tokens and the global features and find that both sIoU and uIoU drop drastically. Finally, we apply the residual feature distillation proposed in [26] and find that though a similar sIoU can be achieved, its uIoU is 9.1% lower than our method.

Ablation studies on the latent embedding generation: In this experiment, we want to clarify the effectiveness of the CLS tokens in the latent embedding generation as shown in Table VII. First, we set the methods without latent embedding as the baseline. Then we use only local CLS tokens to calibrate the latent text embeddings and find that the hIoU improves due to the 6.0% improvements in uIoU. Then, we only use the global CLS tokens, we find that compared with local CLS tokens, the hIoU drops 0.2% due to the uIoU decrease.

Besides, we also conduct experiments on how to generate the latent text embeddings as shown in Table VIII. First, we use the local features $\mathbf{F}_u$ for latent classes directly as the latent text embeddings without any generator. Compared with our method, we find that the performance drops due to the uIoU. Then, we use an MLP as the generator to replace the transformer decoder generator to evaluate if the interaction between the dense features and the CLS tokens is important. We find that compared with using only $\mathbf{F}_u$ the uIoU increases but is still lower than our method due to the lower IoU for unseen classes. Next, we directly apply the local CLS token as the latent text embeddings and find that the hIoU drops drastically to 40.2% from 42.3%, and adding an MLP can slightly increase the performance. Compared with the transformer decoder which combines the merits from the local features $\mathbf{F}_u$ and the CLS token, all other methods achieve sub-optimal performance.

Ablation studies on the token bank size: Table IX presents the ablation study on the impact of token bank size on segmentation performance, measured by harmonic IoU (hIoU), seen IoU (sIoU), and unseen IoU (uIoU). The experiments demonstrate that the inclusion of a token bank significantly improves performance compared to not using a token bank (size 0). The optimal token bank size is found to be 6, achieving the highest hIoU (42.3%), sIoU (41.9%), and uIoU (42.7%). However, increasing the token bank size beyond this optimal point (*e.g.*, size 14) leads to a sub-optimal performance.

TABLE XIII: Experiments on the mask generation.

Classes	unseen	seen	all
mIoU	58.5	53.1	49.5

Ablation studies on the size of K-Means: Table X summarizes the ablation study on the effect of window size **K** in mask generation. When using a single window size, a larger window (size 7) outperforms a smaller one (size 3), achieving 41.8% hIoU, 41.3% sIoU, and 42.3% uIoU. Notably, combining multiple window sizes (3 and 7) yields the best performance across all metrics, with 42.3% hIoU, 41.9% sIoU, and 42.7% uIoU. This result highlights the effectiveness of multi-scale aggregation in capturing complementary spatial information, which enhances the model's segmentation capability.

Ablation studies on feature aggregation in global distillation: Table XI shows an ablation study on feature aggregation strategies, including mean pooling, max pooling, and attention-based aggregation. Mean pooling performs the worst due to over-smoothing, while max pooling improves performance by emphasizing the strongest responses. Attention-based aggregation achieves the best results (42.3% hIoU, 41.9% sIoU, 42.7% uIoU) by dynamically aggregating features, effectively balancing seen and unseen class contributions.

Comparison between other pseudo mask generation and our method: Table XII demonstrates the effectiveness of our pseudo mask generation strategy. Unlike prior methods relying on latent class names [53–55], which violate the zero-shot setting, our approach requires no class-specific information. Compared to Mask Proposal and Panoptic Cut, our method achieves the highest hIoU with comparable GFLOPs, achieving a strong balance between accuracy and efficiency.

C. Accuracy of the generated masks for latent classes

Readers may wonder if the generated masks are accurate enough to serve as the pseudo masks for latent classes. Therefore, we conduct an experiment on the VOC dataset [67], which contains 20 classes. We first split the dataset into seen classes (15 classes) and unseen classes (5 classes). Since the generated masks lack class labels, we convert the ground truth masks into binary masks. During evaluation, the generated mask with the highest IoU is selected as the prediction. The IoU metric is then used to assess the alignment between the selected generated mask and the corresponding ground truth mask. As shown in Table XIII, we evaluate the performance of our method, which notably requires no training and relies solely on clustering. When testing on unseen classes, our approach achieves an impressive 58.5% mIoU. Even for seen classes, the mIoU remains at 53.1%. Finally, when evaluated across all classes, the mIoU reaches 49.5%, demonstrating

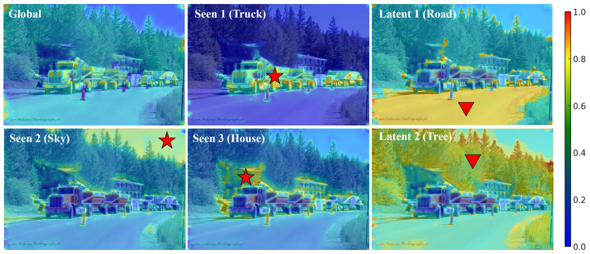


Fig. 5: Similarities between the CLS tokens and dense features, with the red star indicating seen classes and the red triangle indicating latent classes.

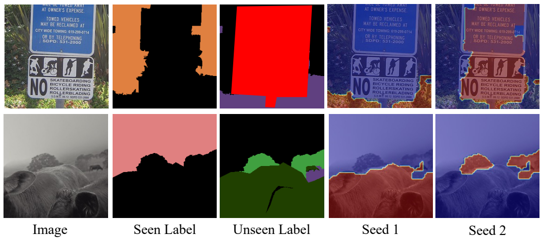


Fig. 6: The visualization of pseudo masks for latent classes. For different images, different classes can be found. The 'seed1' indicates the clustering results for the first seed.

the effectiveness of our clustering-based method in generating high-quality masks without any training or fine-tuning.

D. Qualitative Analysis

The visualization of the similarity between CLS tokens and dense features: We aim to determine whether the distillation process can identify the representative regions. Therefore, we visualize the similarities between the CLS tokens and the dense features as shown in Fig. 5. First, we visualize the similarities between the global CLS tokens and the dense features. We can find that all the areas correspond to the global tokens. Then, we obtain local CLS tokens for the seen areas, *e.g.*, truck (top middle) and house (bottom middle), and we can find that the correspondences are also class-specific. Finally, we generate pseudo masks for the unannotated areas, *i.e.*, road (top right), and tree (bottom right), and calculate their correspondence. We can also achieve the expected results.

The roles of latent class mining: As shown in Fig. 6, this figure highlights the capability of our approach to discover latent classes. Specifically, the results depicted are obtained from two images. Our approach can identify the meaningful objects that are not annotated in the original dataset, demonstrating its latent for discovering unseen or unannotated entities. For instance, in the top of Fig. 6, the latent classes 'tree' and the 'signs' can be found by different seeds. Besides, in the bottom of Fig. 6, the latent classes 'tree' and 'cow' can also be found, indicating our effectiveness.

Qualitative Analysis of Each Module. Fig. 7 shows visual comparisons demonstrate the contribution of each module. Without global distillation, predictions become fragmented and confuse unseen classes (*e.g.*, cow vs. giraffe). Removing local distillation causes semantic inconsistency within objects, while omitting prototype calibration leads to imprecise boundaries. In contrast, our full model produces accurate and consistent segmentation, highlighting the effectiveness of all components.

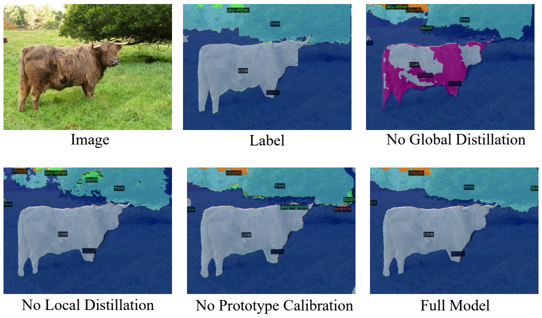


Fig. 7: The qualitative results of each module.

Failure cases: Fig. 9 shows representative failure cases of our method. In the first row, the model misclassifies a concrete as a common wall and confuses bed with carpet, indicating difficulties in fine-grained object recognition under indoor conditions. In the second row, the model fails to correctly segment the tree and leaves. These results suggest that our method still struggles with small object discrimination.

The visualization of prediction: We visualize the prediction of our method as shown in Fig. 8. Compared with SOTA methods, *i.e.*, ZegCLIP [5], our method can obtain exceptional results on both seen and unseen categories. For example, the 'trees' in the fourth image are classified as another unseen class (road) in ZegCLIP. However, our method can correctly recognize it.

E. Discussions

Discussion on CLIP-to-Seg Distillation: For global distillation, our method relies solely on the CLS token representing the entire image, rather than extracting CLS tokens for individual regions [53]. Furthermore, our global distillation adopts a whole-vision distillation approach and performs feature-level aggregation instead of patch-level aggregation, which may suffer from the size of patches and hurt the pixel-level segmentation, compared to [70]. For local distillation, unlike methods that focus exclusively on pulling positive pairs closer [53], our approach also pushes negative pairs from different classes further apart, ensuring more robust class separation.

The primary reason for selecting CLIP as teacher model is the capability to perform vision-language matching, which helps segmentation models generalize to classes that do not appear in training. Besides, while other vision-foundation models, such as SAM [71], excel at delineating object boundaries, they cannot determine whether these segments belong to the same class or tell what class it is. These limitations make other vision-foundation models less suitable for semantic segmentation.

Discussion on Latent Embedding Generation: Readers may be concerned about whether this method violates the zero-shot setting. We argue that our method does not violate the zero-shot setting for the following reasons. First, existing methods such as [5, 11] also incorporate the loss from unannotated areas by pushing the features in these areas away from seen text embeddings. This ensures that features in unannotated regions are less biased towards seen classes and enforces that these features belong to unseen classes during inference. Our approach achieves a similar goal but in a different manner: instead of explicitly enforcing feature separation, we leverage clustering to impose a self-organizing structure on

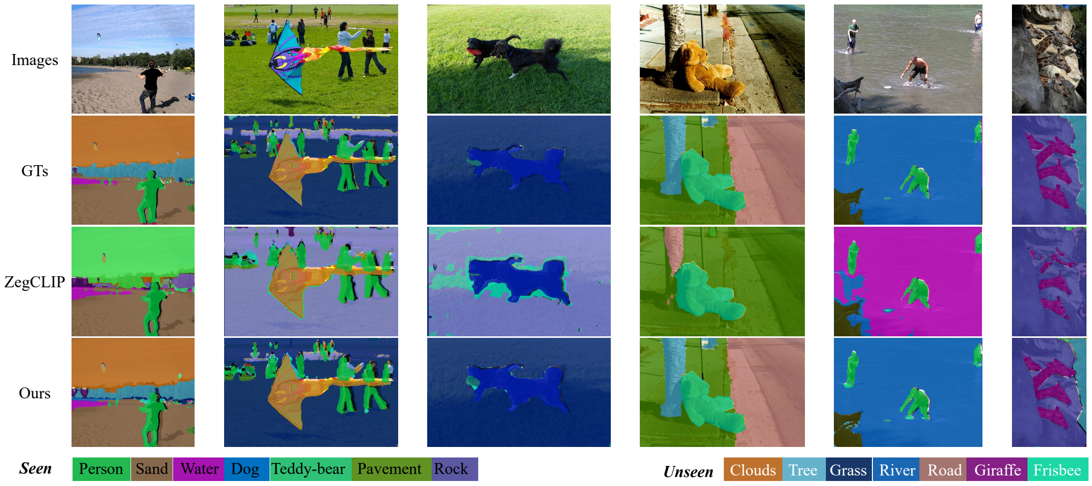


Fig. 8: Visualization comparison between ZegCLIP and our methods.

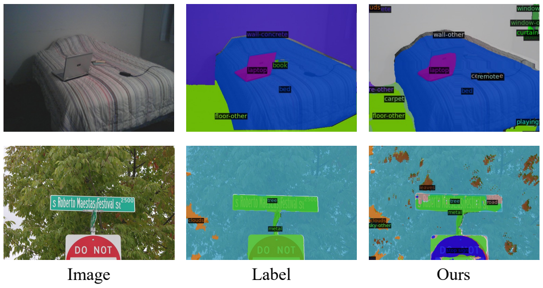


Fig. 9: The qualitative results of failure cases.

unannotated regions. Importantly, this clustering process is entirely unsupervised and does not utilize any labels from unseen classes. The clustering of features is solely driven by their feature similarity, without any guidance from class-level text embeddings. Consequently, the clustering only influences the spatial coherence of features within unannotated regions rather than introducing information that could compromise the zero-shot assumption. Second, we do not use unseen text embeddings during training, and the latent text embeddings merely come from visual features and serve as clustering centers. These clustering centers change dynamically with each image rather than remaining fixed like a classifier. They are only used to group similar features together and do not impose any fixed class labels, ensuring that the model does not learn specific unseen categories during training.

V. Conclusion

In this paper, we propose the CLIP-to-Seg Distillation framework to overcome the limitations of directly adapting CLIP for segmentation tasks. Our approach integrates both global and local distillation strategies to transfer CLIP's zero-shot generalization capabilities to closed-set segmentation models. By aligning dense features from segmentation models with CLS tokens from CLIP at both global and local levels, we facilitate effective distillation from CLIP to pixel-level segmentation models. Additionally, introducing synthesized text embeddings for latent classes enhances the model's ability to generalize to new concepts. Without adding extra parameters or computational overhead, our method achieves state-of-the-art performance on zero-shot segmentation benchmarks, offering a flexible and efficient solution to extend the generalization capabilities of existing segmentation models.

Limitation and Future Works: Though effective, our method still has some drawbacks. The pseudo masks and the text embeddings for latent classes are not accurate enough, leading to sub-optimal performance compared to the fully supervised method. In the future, we aim to produce more accurate pseudo masks and pseudo text embeddings.

Acknowledgment

Support for this work was given by the Toyota Motor Corporation (TMC) and JSPS KAKENHI Grant Number 23K28164 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] X. Zheng, Y. Luo, P. Zhou, and L. Wang, "Distilling efficient vision transformers from cnns for semantic segmentation," *Pattern Recognition*, vol. 158, p. 111029, 2025.

[3] Y. Liu, P. Wu, M. Wang, and J. Liu, "Cpal: Cross-prompting adapter with loras for rgb+ x semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[4] J. Ding, N. Xue, G.-S. Xia, and D. Dai, "Decoupling zero-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 583–11 592.

[5] Z. Zhou, Y. Lei, B. Zhang, L. Liu, and Y. Liu, "Zegclip: Towards adapting clip for zero-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 175–11 185.

[6] 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 *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[7] Y. Zhao, J. Sun, L. Zhang, and H. Lu, "Focusclip: Focusing on anomaly regions by visual-text discrepancies," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.

[8] S. Wu, W. Zhang, L. Xu, S. Jin, X. Li, W. Liu, and C. C. Loy, "CLIPSelf: Vision transformer distills itself for open-vocabulary dense prediction," in *The Twelfth International Conference on Learning Representations*, 2024.

[9] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "Side adapter network for open-vocabulary semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2945–2954.

[10] C. Han, Y. Zhong, D. Li, K. Han, and L. Ma, "Open-vocabulary semantic segmentation with decoupled one-pass network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1086–1096.

[11] Y. Zhang, M.-H. Guo, M. Wang, and S.-M. Hu, "Exploring regional clues in clip for zero-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3270–3280.

[12] Z. Zhang, W. Ke, Y. Zhu, X. Liang, J. Liu, Q. Ye, and T. Zhang, "Language-driven visual consensus for zero-shot semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[13] 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 *European Conference on Computer Vision*. Springer, 2022, pp. 736–753.

[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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 797–807.

[15] Y. Kim, J. Park, Y. Jang, M. Ali, T.-H. Oh, and S.-H. Bae, "Distilling global and local logits with densely connected relations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6290–6300.

[16] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[17] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "Segnext: Rethinking convolutional attention design for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1140–1156, 2022.

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

[19] 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," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[20] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 801–818.

[21] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 764–773.

[22] 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.

[23] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6881–6890.

[24] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1290–1299.

[25] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 864–17 875, 2021.

[26] X. Huang, H. Zhou, K. Yao, and K. Han, "FROSTER: Frozen CLIP is a strong teacher for open-vocabulary action recognition," in *The Twelfth International Conference on Learning Representations*, 2024.

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

[28] Z. Yang, A. Zeng, Z. Li, T. Zhang, C. Yuan, and Y. Li, "From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 185–17 194.

[29] 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.

[30] 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.

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

[32] S. He, H. Ding, and W. Jiang, "Primitive generation and semantic-related alignment for universal zero-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 238–11 247.

[33] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4904–4916.

[34] Y. Zhang, X. Huang, J. Ma, Z. Li, Z. Luo, Y. Xie, Y. Qin, T. Luo, Y. Li, S. Liu *et al.*, "Recognize anything: A strong image tagging model," *arXiv preprint arXiv:2306.03514*, 2023.

[35] R. Pei, J. Liu, W. Li, B. Shao, S. Xu, P. Dai, J. Lu, and Y. Yan, "Clipping: Distilling clip-based models with a student base for video-language retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 983–18 992.

[36] Z. Gu, S. Zhou, L. Niu, Z. Zhao, and L. Zhang, "Context-aware feature generation for zero-shot semantic segmentation," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1921–1929.

[37] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European Conference on Computer Vision*. Springer, 2022, pp. 709–727.

[38] 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 *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.

[39] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," in *International Conference on Learning Representations*, 2022.

[40] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *European Conference on Computer Vision*. Springer, 2022, pp. 540–557.

[41] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *European Conference on Computer Vision*. Springer, 2022, pp. 696–712.

[42] Z. Ding, J. Wang, and Z. Tu, "Open-vocabulary panoptic segmentation with maskclip," *arXiv preprint arXiv:2208.08984*, 2022.

[43] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 446–19 455.

[44] Q. Yu, J. He, X. Deng, X. Shen, and L.-C. Chen, "Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional CLIP," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[45] S. Jiao, Y. Wei, Y. Wang, Y. Zhao, and H. Shi, "Learning mask-aware clip representations for zero-shot segmentation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 35 631–35 653, 2023.

[46] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2955–2966.

[47] X. Li, H. Yuan, W. Li, H. Ding, S. Wu, W. Zhang, Y. Li, K. Chen, and C. C. Loy, "Omg-seg: Is one model good enough for all segmentation?" in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 27 948–27 959.

[48] D. Kang and M. Cho, "In defense of lazy visual grounding for open-vocabulary semantic segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 143–164.

[49] J. Wu, X. Li, S. Xu, H. Yuan, H. Ding, Y. Yang, X. Li, J. Zhang, Y. Tong, X. Jiang *et al.*, "Towards open vocabulary learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 7, pp. 5092–5113, 2024.

[50] F. Li, H. Zhang, P. Sun, X. Zou, S. Liu, C. Li, J. Yang, L. Zhang, and J. Gao, "Segment and recognize anything at any granularity," in *European Conference on Computer Vision*. Springer, 2024, pp. 467–484.

[51] H. Yuan, X. Li, C. Zhou, Y. Li, K. Chen, and C. C. Loy, "Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively," in *European Conference on Computer Vision*. Springer, 2024, pp. 419–437.

[52] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *International Conference on Learning Representations*, 2022.

[53] M. Gao, C. Xing, J. C. Niebles, J. Li, R. Xu, W. Liu, and C. Xiong, "Open vocabulary object detection with pseudo bounding-box labels," in *European Conference on Computer Vision*. Springer, 2022, pp. 266–282.

[54] S. Xu, X. Li, S. Wu, W. Zhang, Y. Tong, and C. C. Loy, "Dst-det: Simple dynamic self-training for open-vocabulary object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[55] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 901–16 911.

[56] Z. Qin, Y. Chen, G. Zhu, E. Zhou, Y. Zhou, Y. Zhou, and C. Zhu, "Enhanced pseudo-label generation with self-supervised training for weakly-supervised semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 7017–7028, 2024.

[57] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742.

[58] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[59] Y. Xian, S. Choudhury, Y. He, B. Schiele, and Z. Akata, “Semantic projection network for zero-and few-label semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8256–8265.

[60] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[61] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez, “Zero-shot semantic segmentation,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[62] J. Cheng, S. Nandi, P. Natarajan, and W. Abd-Almageed, “Sign: Spatial-information incorporated generative network for generalized zero-shot semantic segmentation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9556–9566.

[63] D. Baek, Y. Oh, and B. Ham, “Exploiting a joint embedding space for generalized zero-shot semantic segmentation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9536–9545.

[64] K. Kim, Y. Oh, and J. C. Ye, “Otseg: Multi-prompt sinkhorn attention for zero-shot semantic segmentation,” in *Proceedings of the European Conference on Computer Vision*, 2024.

[65] F. Tung and G. Mori, “Similarity-preserving knowledge distillation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1365–1374.

[66] X. Wang, T. Fu, S. Liao, S. Wang, Z. Lei, and T. Mei, “Exclusivity-consistency regularized knowledge distillation for face recognition,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 325–342.

[67] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge: A retrospective,” *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[68] H. Caesar, J. Uijlings, and V. Ferrari, “Coco-stuff: Thing and stuff classes in context,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1209–1218.

[69] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 891–898.

[70] Z. Ma, G. Luo, J. Gao, L. Li, Y. Chen, S. Wang, C. Zhang, and W. Hu, “Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 074–14 083.

[71] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, “Segment anything,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

JIALEI CHEN (Student Member, IEEE) received the B.Eng. and M.Eng. degrees from Northeastern University, Shenyang, China in 2019 and 2022. He is currently pursuing a Ph.D. degree in information science from Nagoya University, Japan. His main research interests include semantic segmentation, zero-shot learning, and image processing.

Zhenzhen Quan received the B.S. degree from Shandong University of Science and Technology, Qingdao, China, in 2014 and received the M.S. degree from Northeastern University, Shenyang, Liaoning, China, in 2017. She is currently working toward the Ph.D. degree with the School of Information Science and Engineering, Shandong University, Qingdao, China. Her research interests include action recognition, computer vision, and machine learning.

CHENKAI ZHANG received the B.Eng. and B.A. degrees from Dalian University of Technology, Dalian, China in 2019, and B.Eng. and M.Eng. degree from Ritsumeikan University, Shiga, Japan in 2019 and 2022. He is currently pursuing a Ph.D. degree in information science from Nagoya University, Japan. His main research interests include explainable artificial intelligence and the reliability of automatic driving.

Xu Zheng (Student Member, IEEE) is a Ph.D. student in the Visual Learning and Intelligent Systems Lab, Artificial Intelligence Thrust, The Hong Kong University of Science and Technology, Guangzhou Campus (HKUST-GZ). He got his B.E. and M.S. from Northeastern University, China. His research interests include multi-modal learning, sensing and perception techniques.

DAISUKE DEGUCHI (Member, IEEE) received the B.Eng. and M.Eng. degrees in engineering and the Ph.D. degree in information science from Nagoya University, Japan, in 2001, 2003, and 2006, respectively. He became a Postdoctoral Fellow at Nagoya University, in 2006. From 2008 to 2012, he was an Assistant Professor at the Graduate School of Information Science. From 2012 to 2019, he was an Associate Professor at the Information Strategy Office. Since 2020, he has been an Associate Professor with the Graduate School of Informatics. He is working on object detection, segmentation, recognition from videos, and their applications to ITS technologies, such as detection and recognition of traffic signs. He is a member of IEICE and IPS Japan.

HIROSHI MURASE (Life Fellow, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees in electrical engineering from Nagoya University, Japan. In 1980, he joined Nippon Telegraph and Telephone Corporation (NTT). From 1992 to 1993, he was a Visiting Research Scientist with Columbia University, New York. Since 2003, he has been a Professor with Nagoya University. Since 2021, he has been an Emeritus Professor. His research interests include computer vision, pattern recognition, and multimedia information processing. He is a fellow of the IPSJ and the IEICE. He was awarded the IEEE CVPR Best Paper Award, in 1994, the IEEE ICRA Best Video Award, in 1996, the IEICE Achievement Award, in 2002, the IEEE Multimedia Paper Award, in 2004, and the IEICE Distinguished Achievement and Contributions Award, in 2018. He received the Medal with Purple Ribbon from the Government of Japan, in 2012.