Class Token as Proxy: Optimal Transport-assisted Proxy Learning for Weakly Supervised Semantic Segmentation

Supplementary Material

A1. Hyper-parameter Analysis

Analysis of Hyper-parameter K. In Eq. (6), the top-K most related patch tokens are selected to learn proxies alongside their corresponding class token. We perform a hyperparameter analysis for the value of K, as shown in Tab. A1. The best performance is achieved when K equals 5. When K is less than 5, increasing its value incorporates more relevant patch tokens, leading to more comprehensive proxy construction. However, further increasing K beyond 5 results in performance degradation. This could be due to the inclusion of patch tokens that do not belong to the current class, which negatively impacts the final activation performance.

K	1	3	5	7	10
mIoU(%)	72.5	73.1	**73.4**	73.2	72.2

Table A1. Ablation study of top-K selected patch tokens in Eq. (6) on PASCAL VOC 2012 train set. The mIoU (%) metric is used for evaluation.

Analysis of Hyper-parameter λ. In Eq. (12), λ is employed as the weight coefficient for OT-assisted proxy-patch contrastive loss. A hyperparameter analysis for λ is conducted as shown in Tab. A2. When $\lambda = 0$, the $\mathcal{L}_{ppc}$ loss is excluded from training. Increasing λ from 0 to 0.2 improves performance, demonstrating the effectiveness of the proposed $\mathcal{L}_{ppc}$ loss. However, when λ exceeds 0.2, performance declines, indicating that 0.2 is the optimal value for balancing the loss functions.

λ	0	0.1	0.2	0.3	0.5
mIoU(%)	71.5	72.6	**73.4**	73.0	71.9

Table A2. Ablation study of weight coefficient λ in Eq. (12) on PASCAL VOC 2012 train set. The mIoU (%) metric is used for evaluation.

A2. Integration Details on CLIP-ES

CLIP-ES [8] explores leveraging Contrastive Language-Image Pre-training models (CLIP) to localize different categories in WSSS with image-level labels, demonstrating superior performances. Further methods, such as [10, 19], integrate their methods into the CLIP-ES model and achieve significant performances. However, CLIP-ES [8] is a training-free model, and it uses the vanilla Vision Transformer (ViT) structure, which only integrates a class token to classify an image. In this section, we illustrate the procedure of integrating our framework into the CLIP-ES [8].

Prototype Generation. To adapt the proposed OT-assisted proxy learning strategy to structures without multiple class tokens, class prototypes are generated and applied to the framework for seamless integration. Specifically, the prototypes $\mathbf{Q} \in \mathbb{R}^{(C+1)\times D}$ are generated using the pseudo labels from the classifier CAM $\mathcal{M}^{\text{cls}}$ as follows:

$$\mathbf{Q} = MAP\big(Mask\left(\mathcal{M}^{\text{cls}}\right) \odot \mathbf{F}\big), \tag{13}$$

where $MAP(\cdot)$ denotes average pooling over the spatial dimensions of the flattened features. The function $Mask(\cdot)$ generates a binary mask Mask $\in \mathbb{R}^{HW\times(C+1)}$ based on the pseudo label results, where $\text{Mask}_{i,c} = 1$ indicates that the feature at position i is assigned to class c. Here, $\mathbf{F}$ represents the feature map produced by the encoder, and $\odot$ denotes element-wise multiplication.

These prototypes capture the discriminative characteristics of each class, as they are constructed from features with confident CAM scores. Consequently, they effectively function as 'class tokens'. Subsequently, the proposed two learning strategies are applied to these prototypes.

OT-assisted proxy learning. The first strategy aims to learn proxies that preserve classification capabilities while capturing essential foreground characteristics, leading to more comprehensive and accurate resulting CAMs. Prototypes and encoded features are modeled as two distinct distributions. The objective function is formulated as:

$$\begin{aligned} &\mathbf{T}^*(\mathbf{S}) = \arg\max_{\mathbf{T}} \sum_{i=1}^{M}\sum_{j=1}^{N} \mathbf{T}_{i,j}\mathbf{S}_{i,j}, \\ &\text{s. t. } \mathbf{T}\mathbf{1}_N = \frac{1}{M}\mathbf{1}_M,\ \mathbf{T}^T\mathbf{1}_M = \mathbf{x}, \end{aligned} \tag{14}$$

where $\mathbf{S} \in \mathbb{R}^{M\times N}$ is a similarity map and the element $\mathbf{S}_{i,j}$ represents the cosine similarity between the i-th feature and the j-th prototype. N is the total number of prototypes and $N = (C+1)$. M is the number of features and $M = HW$. $\mathbf{1}_M$ is a column vector of ones with the dimension M. The marginal constraint $\mathbf{x}$ from Eq. (4) is also leveraged in this structure.

Once the optimization process is completed, the subset of features is selected based on the optimized transport plan scores, which quantify their relevance to the prototype:

$$\mathcal{H}_j = \left\{ \mathbf{F}_i \mid i \in \arg\max_i \left(\mathbf{T}^*_{i,j},\ \text{top} = K\right) \right\}, \tag{15}$$

where $\mathcal{H}_j$ represents the top-K featuress selected for the j-th prototype.

Subsequently, prototype $\mathbf{Q}_c$ and the corresponding features in set $\mathcal{H}_c$ are concatenated together. The concatenated result is then passed through a convolutional layer to generate a proxy $\mathcal{P}_c$ for class c:

$$\mathcal{P}_c = \text{Conv}(\text{Concat}[\mathbf{Q}_c, \mathcal{H}_c]), \tag{16}$$

where $\mathcal{P} \in \mathbb{R}^{(C+1)\times D}$ are learned proxies. $\text{Conv}(\cdot)$ denotes a convolutional layer, and $\text{Concat}[\cdot]$ represents the concatenation operation.

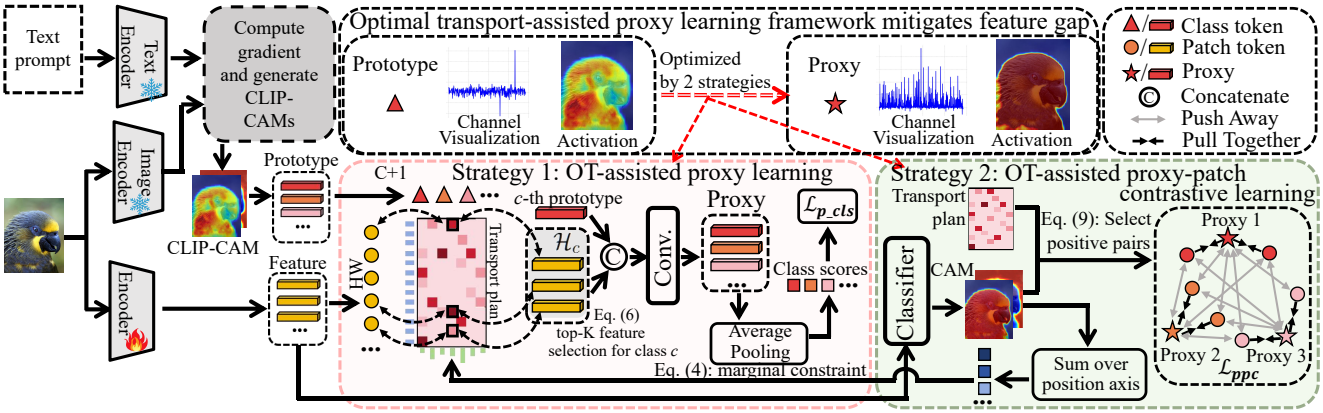

Figure A1. Illustration of the OTPL framework on CLIP-ES [8]. Prototypes often fail to fully activate foreground features due to the gap between prototypes and features. Our framework learns proxies to bridge this gap through two strategies. Strategy 1: Optimal Transport-assisted proxy learning to construct proxies. Image and text prompts are separately fed into the frozen CLIP encoders to extract their respective features. The features are then leveraged to generate initial GradCAMs [9], using the gradients of calculating cosine similarity between them. These initial GradCAMs are subsequently refined using Sinkhorn normalization following previous works [8, 17], resulting in CLIP-CAMs. Confident CAM results from CLIP-CAMs are used to construct prototypes. These prototypes and encoded features are leveraged to construct a cost matrix for Optimal Transport to conduct an optimization. This process is guided by a marginal constraint derived from CAM predictions to emphasize class importance. The resulting transport plan provides a probability distribution for assigning features to prototypes. By combining a prototype with its most related top-K features based on the transport plan, a proxy is learned that not only maintains the classification ability but also captures essential foreground characteristics, thus better activating relevant features. Strategy 2: OT-assisted prototype-feature contrastive learning to further align the generated proxies with confident features. The CAM predictions and optimized transport plan jointly identify positive pairs for each proxy, refining proxy separability and their capacity to capture non-discriminative foreground characteristics through contrastive training. For simplicity, the proxy generation process is illustrated with a specific class.

To preserve the classification capabilities of the generated proxies, supervised learning is applied to these proxies using image-level labels with a multi-label soft margin classification loss $\mathcal{L}_{p\_cls}$, following previous methods [14, 15]. This can be defined as follows:

$$\begin{aligned}\mathcal{L}_{p\_cls} = \frac{1}{|\mathcal{C}|}\sum_{c=1}^{|\mathcal{C}|} l_c \log\left(\text{sigmoid}\left(AP\left(\mathcal{P}_c\right)\right)\right) \\ + \left(1-l_c\right)\log\left(1-\text{sigmoid}\left(AP\left(\mathcal{P}_c\right)\right)\right),\end{aligned} \tag{17}$$

where $AP(\cdot)$ denotes the average pooling function, l_c represents the image-level label for the c-th class, and $\mathcal{C}$ is the set of foreground classes.

OT-assisted proxy-patch contrastive learning. OT-assisted proxy-patch contrastive learning strategy is proposed to better align the previously generated proxies with confident features. In our framework, positive samples are identified through the joint analysis of CAM $\mathcal{M}^{\text{cls}}$ and the optimized transport plan $\mathbf{T}^*$ from previous step, which can be mathematically defined as:

$$\mathcal{R}_c^+ = \left\{ \mathbf{F}_i \mid \arg\max_j \mathcal{M}^{\text{cls}}_{i,j} = c \text{ and } \arg\max_k \mathbf{T}^*_{i,k} = c \right\}, \tag{18}$$

where j and k are class indexes, and i is the position index. The OT-assisted proxy-patch contrastive learning loss is then formulated as:

$$\mathcal{L}_{ppc} = -\frac{1}{|\mathcal{R}_c^+|}\sum_{c\in\mathcal{C}}\sum_{\mathbf{F}_i^+\in\mathcal{R}_c^+}\log\frac{\exp(\mathcal{P}_c\cdot\mathbf{F}_i^+/\tau)}{\sum_{\mathbf{F}_i\in\mathcal{R}_c}\exp(\mathcal{P}_c\cdot\mathbf{F}_i/\tau)}, \tag{19}$$

where $|\mathcal{R}_c^+|$ represents the number of positive pairs within the prototypes and the features, $\mathcal{R}_c$ is the set of all features, and τ is the temperature factor controlling the sharpness of the contrastive loss.

Training Details on CLIP-ES. In our experiment for obtaining prototype CAMs on CLIP-ES [8], we train our model on a single NVIDIA RTX 3090 GPU with 24GB memory, using a batch size of 16. To ensure robust and consistent results, we adopt the same data augmentation strategies as previous works, including random flipping, random scaling, and cropping, as described in [3, 6]. For pseudo label generation in the PASCAL VOC 2012 dataset [4], we utilize the IRN [1] post-processing method to refine the CAMs. However, due to the computational cost, we directly use DenseCRF [5] as the post-processing method in the MS COCO 2014 dataset [7] following SIPE [3] and SFC [18]. For further segmentation model training, we employ ResNet101-based DeepLabV2 [2] and follow the settings established by previous methods [11, 12].

A3. Additional Visualization Results

Our analysis reveals that classification-focused class tokens naturally exhibit partial activation patterns. Here, we provide more

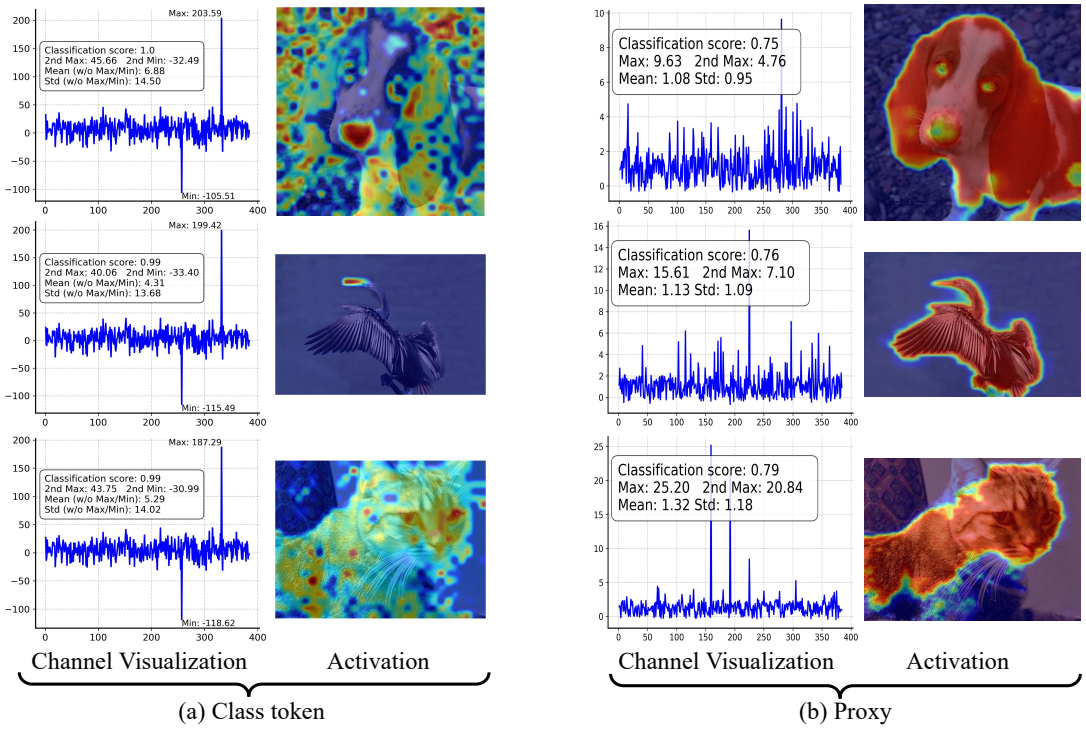

Figure A2. Visualization of channels and activation results (I).

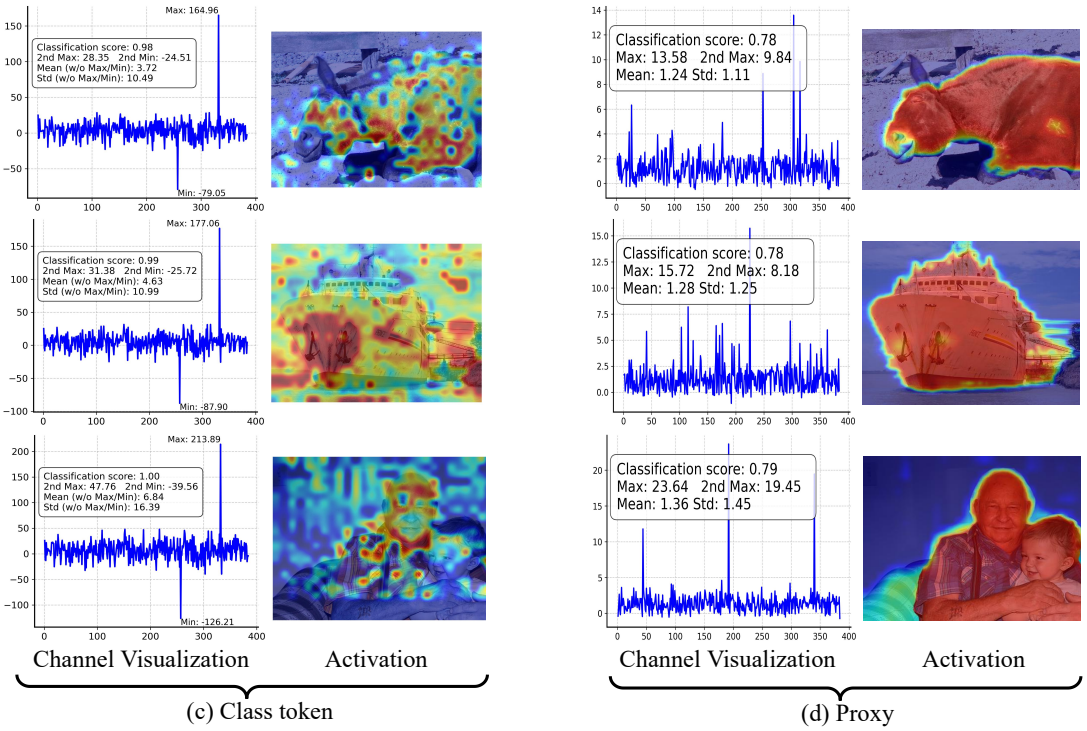

Figure A3. Visualization of channels and activation results (II).

visualization comparison results between class tokens and our proxies, as shown in Fig. A2 and Fig. A3. As visualized in these figures, class tokens predominantly activate discriminative foreground regions through specific encoded channels, leav-

ing other semantically relevant areas under-activated – a phenomenon systematically analyzed in Sect. 3. While previous approaches [14, 15] attempt to address this limitation through computationally intensive patch-to-patch attention mechanisms in ViT, the results are not satisfactory enough for all images. In contrast, our method introduces a more efficient solution: the proposed proxies directly activate comprehensive feature representations, achieving substantially more complete foreground coverage, as evidenced in Fig. A2 (b) and Fig. A3 (d).

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, pages 2209–2218, 2019. 7, 2

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2017. 1, 7, 2

[3] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *CVPR*, pages 4288–4298, 2022. 2, 6, 7

[4] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 84(4):98–136, 2015. 6, 2

[5] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, pages 834–848, 2011. 7, 2

[6] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *CVPR*, pages 4071–4080, 2021. 6, 7, 2

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, page 740–755, 2014. 6, 2

[8] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *CVPR*, pages 15305–15314, 2023. 6, 7, 8, 1, 2

[9] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 6, 2

[10] Feilong Tang, Zhongxing Xu, Zhaojun Qu, Wei Feng, Xingjian Jiang, and Zongyuan Ge. Hunting attributes: Context prototype-aware learning for weakly supervised semantic segmentation. In *CVPR*, pages 3324–3334, 2024. 2, 6, 7, 1

[11] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *CVPR*, pages 12275–12284, 2020. 2

[12] Yuanchen Wu, Xiaoqiang Li, Songmin Dai, Jide Li, Tong Liu, and Shaorong Xie. Hierarchical semantic contrast for weakly supervised semantic segmentation. In *IJCAI*, pages 1542–1550, 2023. 2

[13] Yuanchen Wu, Xichen Ye, Kequan Yang, Jide Li, and Xiaoqiang Li. Dupl: Dual student with trustworthy progressive learning for robust weakly supervised semantic segmentation. In *CVPR*, pages 3534–3543, 2024. 7

[14] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *CVPR*, 2022. 1, 2, 5, 6, 7, 8, 4

[15] Lian Xu, Mohammed Bennamoun, Farid Boussaid, Hamid Laga, Wanli Ouyang, and Dan Xu. Mctformer+: Multi-class token transformer for weakly supervised semantic segmentation. *IEEE TPAMI*, 2024. 1, 2, 3, 5, 6, 7, 8, 4

[16] Sung-Hoon Yoon, Hoyong Kwon, Hyeonseong Kim, and Kuk-Jin Yoon. Class tokens infusion for weakly supervised semantic segmentation. In *CVPR*, pages 3595–3605, 2024. 6, 7

[17] Bingfeng Zhang, Siyue Yu, Yunchao Wei, Yao Zhao, and Jimin Xiao. Frozen clip: A strong backbone for weakly supervised semantic segmentation. In *CVPR*, pages 3796–3806, 2024. 2

[18] Xinqiao Zhao, Feilong Tang, Xiaoyang Wang, and Jimin Xiao. Sfc: Shared feature calibration in weakly supervised semantic segmentation. In *AAAI*, pages 7525–7533, 2024. 7, 2

[19] Xinqiao Zhao, Ziqian Yang, Tianhong Dai, Bingfeng Zhang, and Jimin Xiao. Psdpm: Prototype-based secondary discriminative pixels mining for weakly supervised semantic segmentation. In *CVPR*, pages 3437–3446, 2024. 6, 7, 1