Noname manuscript No.
(will be inserted by the editor)

WeakCLIP: Adapting CLIP for Weakly-supervised Semantic Segmentation

Lianghui Zhu[1] · **Xinggang Wang**[1*] · **Jiapei Feng**[1] · **Tianheng Cheng**[1] · **Yingyue Li**[1] · **Bo Jiang**[1] · **Dingwen Zhang**[2] · **Junwei Han**[2]

Received: date / Accepted: date

Abstract Contrastive language and image pre-training (CLIP) achieves great success in various computer vision tasks and also presents an opportune avenue for enhancing weakly-supervised image understanding with its large-scale pre-trained knowledge. As an effective way to reduce the reliance on pixel-level human-annotated labels, weakly-supervised semantic segmentation (WSSS) aims to refine the class activation map (CAM) and produce high-quality pseudo masks. Weakly-supervised semantic segmentation (WSSS) aims to refine the class activation map (CAM) as pseudo masks, but heavily relies on inductive biases like hand-crafted priors and digital image processing methods. For the vision-language pre-trained model, i.e. CLIP, we propose a novel text-to-pixel matching paradigm for WSSS. However, directly applying CLIP to WSSS is challenging due to three critical problems: 1) the task gap between contrastive pre-training and WSSS CAM refinement, 2) lacking text-to-pixel modeling to fully utilize the pre-trained knowledge, and 3) the insufficient details owning to the $\frac{1}{16}$ down-sampling resolution of ViT. Thus, we propose WeakCLIP to address the problems and leverage the pre-trained knowledge from CLIP to WSSS. Specifically, we first address the task gap by proposing a visual adapter and learnable prompts to extract WSSS-specific representation. We then design a co-attention matching module to model text-to-pixel relationships. Finally, the pyramid adapter and text-guided decoder are introduced to gather multi-level information and integrate it with text guidance hierarchically. WeakCLIP provides an effective and parameter-efficient way to transfer CLIP knowledge to refine CAM. Extensive experiments demonstrate that WeakCLIP achieves the *state-of-the-art* WSSS performance on standard benchmarks, *i.e.*, 74.0% mIoU on the val set of PASCAL VOC 2012 and 46.1% mIoU on the val set of COCO 2014. The source code and model checkpoints are released at https://github.com/hustvl/WeakCLIP.

Keywords Semantic Segmentation · Weakly-supervised Learning · CAM Refinement · CLIP

1 Introduction

Weakly-supervised semantic segmentation (WSSS) is an important task to reduce the pixel-level annotation burden which leverages weak supervisions such as image-level classification labels (Ahn and Kwak, 2018; Kolesnikov and Lampert, 2016; Pathak et al., 2015; Pinheiro and Collobert, 2015; Wang et al., 2020; Wei et al., 2017, 2018; Huang et al., 2018; Jiang et al., 2019; Lee et al., 2019a), points (Bearman et al., 2016), scribbles (Lin et al., 2016; Tang et al., 2018; Vernaza and Chandraker, 2017), and bounding boxes (Dai

Lianghui Zhu
E-mail: lhzh@hust.edu.cn

Xinggang Wang
E-mail: xgwang@hust.edu.cn

Jiapei Feng
E-mail: fjp@hust.edu.cn

Tianheng Cheng
E-mail: thch@hust.edu.cn

Yingyue Li
E-mail: yingyueli@hust.edu.cn

Bo Jiang
E-mail: bjiang@hust.edu.cn

Dingwen Zhang
E-mail: zhangdingwen2006yyy@gmail.com

Junwei Han
E-mail: junweihan2010@gmail.com

[1] Huazhong University of Science and Technology, Wuhan, China
[2] Northwestern Polytechnical University, Xi'an, China
[*] Corresponding Author

et al., 2015; Khoreva et al., 2017; Xu et al., 2015; Song et al., 2019), to generate pseudo pixel-wise segmentation. Among these weak supervisions, the most challenging one is WSSS with only image-level labels. Image-level WSSS methods typically require a class activation map (CAM) (Zhou et al., 2016) to coarsely localize an object. As CAMs generated from Deep Neural Networks (DNNs) are often noisy and prone to errors, many methods (Huang et al., 2018; Feng et al., 2021; Ahn and Kwak, 2018; Ahn et al., 2019) are proposed to refine the CAM using hand-crafted priors (i.e., Random Walk (Ahn and Kwak, 2018)) or improved digital image processing (DIP) algorithms (i.e., Seed Region Growing (Huang et al., 2018)). Refined CAM typically exhibits more accurate semantic information compared to the original CAM. As a result, it can serve as a valuable pseudo mask for training fully-supervised semantic segmentation networks. However, these methods heavily rely on inductive biases from priors and improved DIP algorithms, which will limit their performance and robustness. In this paper, we investigate a CAM refinement method based on large-scale pre-trained foundation models.

The work of CLIP (Radford et al., 2021), demonstrates that the successful approach of task-agnostic large-scale pre-training in Natural Language Processing (NLP) can be effectively transferred to Computer Vision (CV) tasks. There are some studies that have explored the CLIP for weakly-supervised semantic segmentation (WSSS). CLIMS (Xie et al., 2022) employs CLIP-based loss functions as regularization. And CLIP-ES (Lin et al., 2022) utilizes well-designed text prompts and GradCAM (Selvaraju et al., 2017) to enhance the quality of the class activation map (CAM). Different from these methods, we propose to transfer CLIP pre-trained knowledge to WSSS in a text-to-pixel matching paradigm. With the exploration of this novel paradigm, we find that there are three primary challenges in leveraging CLIP for WSSS. 1) **Task gap**: the pre-training objective of CLIP causes the vision encoder to focus more on image-level representation, which is inconsistent with the pixel-level understanding required in WSSS. 2) **Lacking text-to-pixel modeling**: both CLIP and previous CLIP-based WSSS methods lack explicit text-to-pixel modeling, which is crucial for effectively transferring pre-trained knowledge to WSSS. 3) **Insufficient details**: the CLIP-ViT model, with its $\frac{1}{16}$ downsampling resolution, fails to provide sufficient spatial information, particularly fine details, thus limiting the quality of CAM refinement.

To address the challenges mentioned above, we propose WeakCLIP, a novel **Weak**ly-supervised Semantic Segmentation method that leverages the parameter-efficient **CLIP** framework. In WeakCLIP, we tackle the task gap issue by proposing the visual adapter and the learnable language prompt to efficiently learn WSSS-specific visual and text representations. Additionally, we introduce a text-to-pixel co-attention matching module to learn contextually informative pixel and text representations, facilitating text-to-pixel modeling for utilizing CLIP pre-trained knowledge. Moreover, we propose a pyramid adapter and text-guided decoder to enhance the level of details and decode features with text-guidance hierarchically. The proposed WeakCLIP primarily utilizes scalable, large-scale, pre-trained vision-language models. It can be enhanced further when combined with more advanced models. Fig. 1 illustrates the distinctiveness of the WeakCLIP scheme for WSSS compared to previous methods that rely on ImageNet pre-trained models for backbone initialization. The proposed WeakCLIP leverages the knowledge that exists in large-scale pre-trained vision-language models into WSSS through text-pixel matching. Experimental results on PASCAL VOC 2012 (Everingham et al., 2010) and COCO 2014 (Lin et al., 2014) demonstrate that the visual-language features extracted by WeakCLIP exhibit superior semantic accuracy compared to deep visual features used in previous WSSS methods.

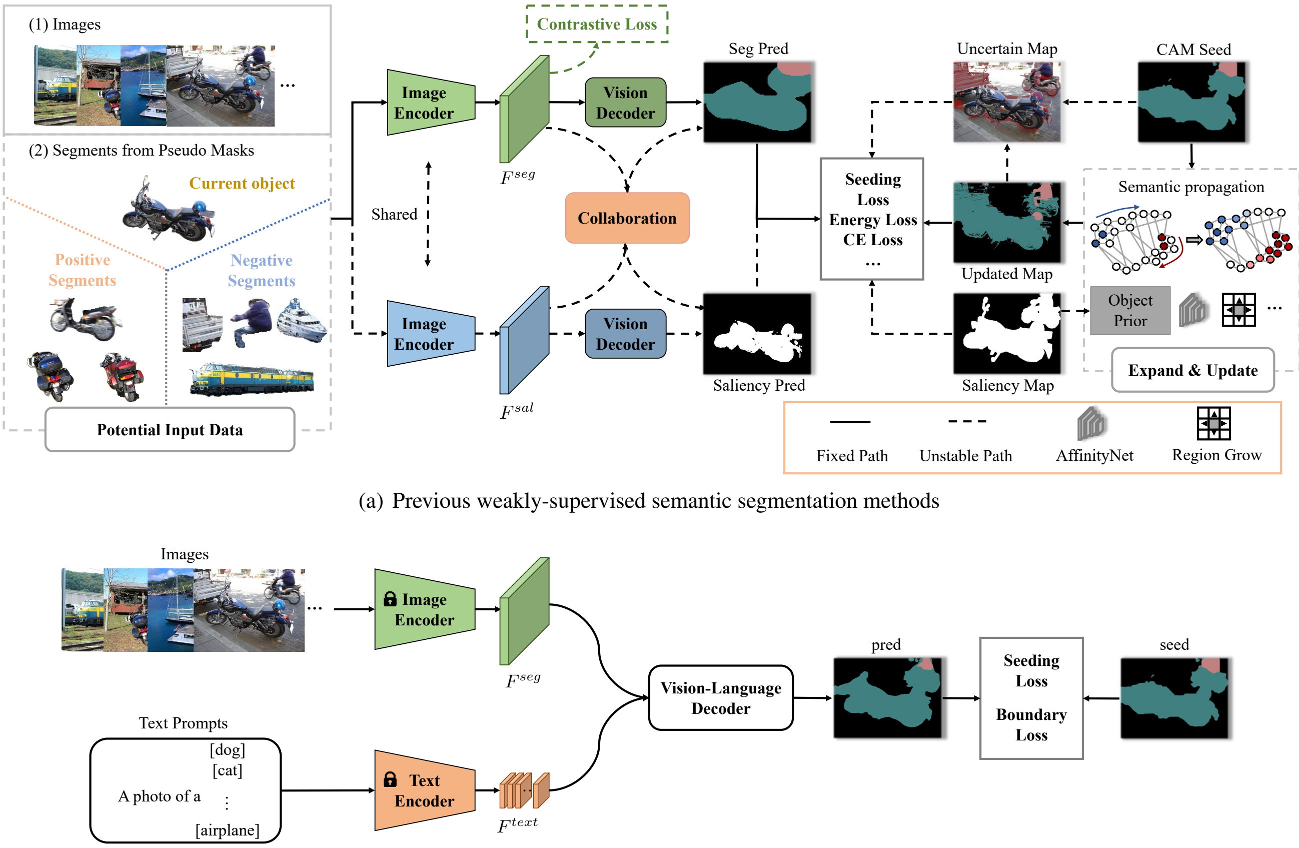
The main contributions of this paper can be summarized as follows:

- We present WeakCLIP, a novel approach that transforms the weakly-supervised semantic segmentation paradigm into a continuous text-to-pixel matching problem. This transformation allows WeakCLIP to utilize natural language guidance from large-scale pre-trained models, reducing the dependence on hand-crafted priors.
- WeakCLIP proposes four key components: efficient learnable prompts, a pyramid adapter, text-pixel co-attention matching, and a text-guided decoder to address the three major challenges encountered in applying CLIP to WSSS, namely task gap, lacking text-to-pixel modeling, and insufficient details.
- Extensive experiments demonstrate that WeakCLIP is both effective and efficient. Through training only 14% parameters, WeakCLIP surpasses the *state-of-the-art* WSSS method on standard benchmarks, with 74.0% mIoU on PASCAL VOC 2012 (Everingham et al., 2010) and 46.1% on COCO 2014 (Lin et al., 2014) validation sets, respectively.

2 Related Work

2.1 Weakly-supervised Semantic Segmentation

In recent years, deep learning has witnessed remarkable progress in semantic segmentation. However, the process of annotating pixel-level ground truth for semantic segmentation is time-consuming, limiting the practical application development. To alleviate this burden, weakly-supervised learning has emerged as an alternative annotation format that requires weaker-level supervision. Numerous weakly-



(a) Previous weakly-supervised semantic segmentation methods

(b) The proposed WeakCLIP scheme

Fig. 1 The previous weakly-supervised semantic segmentation (WSSS) methods tend to have special designs for input, network, loss function, supervision information, semantic expansion, etc. First of all, some methods choose to mine the semantic relationship between objects (Du et al., 2022) or introduce web data (Shen et al., 2017; Wei et al., 2016) in terms of input. Secondly, in terms of network design, some approaches deployed dual backbones with shared weights for multi-task training, such as the concurrent execution of WSSS and saliency prediction (Xu et al., 2021a; Du et al., 2022). Then, in the selection of the loss function, in addition to the commonly used seeding loss (Huang et al., 2018) and energy loss (Zhang et al., 2020a), the recently popular contrastive loss is applied (Du et al., 2022). In terms of supervision information, some methods introduce saliency supervision information (Lee et al., 2021d; Xu et al., 2021a; Yao et al., 2021) or the uncertainty of semantic information (Li et al., 2021a). Finally, some use the additionally introduced location conspicuous (Zhang et al., 2020b), AffinityNet (Ahn and Kwak, 2018), or other traditional regional expansion methods (Huang et al., 2018; Feng et al., 2021) for semantic expansion. As a comparison, the proposed WeakCLIP is designed around scalable pre-trained weights, making it more promising than WSSS methods that rely on hand-crafted priors. Furthermore, WeakCLIP utilizes a novel paradigm of text-to-pixel matching, which efficiently transfers knowledge from pre-trained CLIP.

supervised semantic segmentation algorithms with promising performance have been proposed. These approaches can be categorized based on the types of annotations used, including bounding box-based methods (Dai et al., 2015; Khoreva et al., 2017; Xu et al., 2015; Song et al., 2019), scribble-based methods (Lin et al., 2016; Tang et al., 2018; Vernaza and Chandraker, 2017), point-based methods (Bearman et al., 2016), and image-level label-based methods (Ahn and Kwak, 2018; Kolesnikov and Lampert, 2016; Pathak et al., 2015; Pinheiro and Collobert, 2015; Wang et al., 2020; Wei et al., 2017, 2018; Huang et al., 2018; Jiang et al., 2019; Lee et al., 2019a).

In this paper, we focus on weakly-supervised semantic segmentation using image-level labels. This direction is widely recognized as one of the most challenging aspects in the field. Through the proposed WeakCLIP, high-quality pseudo masks can be attained for the fully-supervised semantic segmentation retraining.

2.2 Image-level Supervised Learning

Image-level labels are one of the most challenging and cost-effective forms of annotation widely utilized in weakly-supervised semantic segmentation (WSSS). However, generating accurate pseudo masks becomes challenging since image-level labels do not provide explicit object localization information. The class activation map (CAM) method proposed by Zhou et al. (Zhou et al., 2016) is commonly used to detect discriminative object regions and generate localization maps known as "seeds." Nonetheless, naive CAM seeds often overlook significant parts of the objects, making them unsuitable as direct proxy ground truth.

To alleviate this problem, various techniques have been proposed. AE-PSL (Wei et al., 2017) introduces an iterative erasing technique that updates regions by erasing previously computed pseudo masks in the raw image. MDC (Wei et al., 2018) suggests using multiple layers with different dilation rates to expand activated regions. SEC (Kolesnikov and Lampert, 2016) proposes a pipeline that incorporates an expansion loss and a CRF constraint loss with the original segmentation loss. DSRG (Huang et al., 2018) presents a seeds region expanding algorithm to gradually enlarge the initial seeds. FickleNet (Lee et al., 2019a) generates and combines diverse activation maps using random connections. OAA (Jiang et al., 2019) accumulates activation maps and trains the CAM to achieve more complete estimations. AffinityNet (Ahn and Kwak, 2018) employs CAM seeds as proxy labels to train an affinity network and utilizes random walks for region expansion. Building upon AffinityNet, IRNet (Ahn et al., 2019) incorporates prior knowledge that boundaries exist along the line between two pixels with different categories. CONTA (Zhang et al., 2020b) employs causal inference to improve the quality of CAM seeds. PMM (Li et al., 2021b) and URN (Li et al., 2021a) employ various strategies to suppress noise during CAM generation. Additionally, co-segmentation (Shen et al., 2017) and STC (Wei et al., 2016) leverage web images to estimate reliable pseudo masks. Some of these approaches rely on hand-crafted rules and carefully adjusted hyperparameters to generate better seeds, while others (Lee et al., 2019b; Yang et al., 2018) utilize additional web data.

In contrast, we propose an effective approach that adapts CLIP to transfer pre-trained knowledge as text guidance. The proposed WeakCLIP contains less inductive bias and exhibits promising potential for dealing with the challenges in WSSS.

2.3 Large-scale Vision-Language Models

For a considerable period of time, pre-trained models on large-scale datasets such as JFT (Sun et al., 2017) and Kinetics (Carreira and Zisserman, 2017) have been extensively utilized across various domains. In recent times, an increasing number of pre-trained models have become available, including those based on supervised learning (Dosovitskiy et al., 2020; He et al., 2016) and self-supervised learning (Caron et al., 2021; Chen et al., 2020b; He et al., 2020).

Furthermore, vision-language pre-training has garnered significant attention (Lei et al., 2021; Lu et al., 2019; Su et al., 2019). Notably, Radford et al. introduced CLIP (Radford et al., 2021), a large-scale pre-trained model trained using contrastive learning on a vast collection of image-text pairs. CLIP demonstrates remarkable transferability across 30 classification datasets. Subsequently, several extensions of CLIP have been proposed to improve the classification ability of CLIP (e.g., CoOp (Zhou et al., 2021), CLIP-Adapter (Gao et al., 2021), Tip-adapter (Zhang et al., 2021b)). Additionally, there are methods that adapt CLIP for different domains. For instance, DenseCLIP (Rao et al., 2021) leverages CLIP to enable dense predictions for fully-supervised object detection and semantic segmentation tasks. On the other hand, MaskCLIP achieves remarkable results in annotation-free segmentation but exhibits limited performance on challenging benchmarks such as COCO 2014 (Lin et al., 2014).

In the WSSS domain, CLIMS (Xie et al., 2022) incorporates CLIP to introduce a set of auxiliary losses, aiding the CAM network in distinguishing background from foreground regions. CLIP-ES (Lin et al., 2022) adopts well-designed text prompts and GradCAM (Selvaraju et al., 2017) to enhance the quality of the class activation map (CAM). However, these methods only utilize the text-to-image matching of CLIP, which could not bring a precise understanding at the pixel level.

Our work proposes a new text-to-pixel paradigm for WSSS that utilizes CLIP pre-trained knowledge in a fine-grained way. Furthermore, we identify the three key challenges of applying CLIP to WSSS. In light of these challenges, we propose WeakCLIP, a novel framework designed to address those issues. By incorporating CLIP into the WSSS framework, WeakCLIP aims to explore a promising WSSS method that could benefit from the advances in large-scale pre-trained models.

3 Method

To better describe the proposed method, we begin by reviewing the methodologies employed in the original CLIP framework (Radford et al., 2021). Subsequently, we present the paradigm transformation of the weakly-supervised semantic segmentation task. Then, we delve into the architecture of our proposed framework, WeakCLIP. Finally, we detail the process of generating high-quality proxy ground truth (PGT) and outline the subsequent retraining procedure.

3.1 Preliminaries: Overview of CLIP

CLIP aligns images and text in semantic space using a contrastive loss. Its encoder combines a transformer for text and either a ResNet or ViT for images. Notably, the CLIP-ViT variant has attracted growing interest among researchers because of its scalability. In the context of our study, we specifically investigate the potential of CLIP-ViT within the weakly-supervised semantic segmentation (WSSS) domain, aiming to enhance the quality and accuracy of pseudo-labels, particularly in the case of coarse class activation map (CAM).

CLIP utilizes 400 million image-text pairs for pre-training and demonstrates strong knowledge transferability. For downstream classification tasks, CLIP generates representations for both textual categories and images, enabling zero-shot classification by computing the similarity between the category and image representations. This approach of computing text and image similarity provides a generalizable method for transferring knowledge from pre-trained models to specific tasks. Previous studies utilized the text-to-image matching ability of CLIP to boost WSSS performance. For example, CLIMS (Xie et al., 2022) incorporates CLIP into IRNet, introducing additional losses and enhancing CAM quality. Similarly, CLIP-ES (Lin et al., 2022) merges CLIP with GradCAM (Selvaraju et al., 2017) and utilizes a specific prompt to achieve an improved CAM. These works lead to an interesting question: Can CLIP's matching capability be leveraged pixel by pixel to enhance pseudo-labels as a highly generalizable semantic inference method?

Addressing this question poses several challenges. Firstly, how to leverage CLIP through text-to-pixel matching is an under-explored area for WSSS. While text-to-pixel matching approaches achieved promising results, we argue that the text-to-pixel matching approach, akin to the pre-training objective, offers a more direct, effective, and robust method for transferring knowledge from CLIP. Secondly, applying CLIP to WSSS presents challenges such as the task gap, lacking text-to-pixel modeling, and insufficient details. Consequently, leveraging CLIP through text-to-pixel matching for WSSS is crucial but challenging, which offers a promising way to improve the quality of pseudo labels.

3.2 Text-to-pixel Matching Paradigm for WSSS

The proposed text-to-pixel matching paradigm differs significantly from previous CLIP-based WSSS approaches (Lin et al., 2022; Xie et al., 2022). These previous methods only deploy the text-to-image matching that lacks the pixel-level fine-grained representation. To solve this problem, we propose text-to-pixel matching to query similarities at pixel level.

Given I as an input image, we first extract its multi-layer feature maps using the image encoder. We employ the CLIP pre-trained ViT-B network as the backbone and consider its multi-layer outputs as the features of I. For the multi-layers of the transformer, we divide them into stages, with each stage encompassing three transformer layers. We represent the feature maps from each stage as f_i, where $i \in \{1, 2, 3, 4\}$ and the shape of f_i is $C \times H \times W$. The size of f_i remains consistent from 1 to 4. Then, CLIP incorporates a projection layer after the last transformer stage, projecting the embeddings into the projected dimension D.

$$\{f_i\} = \Phi_{\text{image}}\left(I\right), i \in \{1, \ldots 4\} \tag{1}$$

$$f_p = \mathrm{Proj}(f_4), \tag{2}$$

$$f_{p,cls} = \mathrm{Proj}(f_{4,cls}). \tag{3}$$

Here, Φ_{image} represents the CLIP ViT-B image encoder, and i denotes the stage index of the encoder. Proj represents the projection layer, and f_p corresponds to the output of the projection layer. We refer to the class token as $f_{4,cls}$, and the projected class token as $f_{p,cls}$. Original CLIP utilizes $f_{p,cls}$ as the whole image representation, and calculates the cosine similarity with the projected text representation t_p for the classification task.

Through our experiments, we discover two important properties of the projected embeddings f_p generated by the CLIP image encoder. Firstly, as the projected embeddings of f_4, f_p retains sufficient spatial information to serve as the feature maps. Secondly, due to the symmetry of the self-attention layer with respect to each input element, f_p exhibits similar characteristics to the classification representation $f_{p,cls}$, as they are both mapped to the same semantic space.

Building upon these observations, we define the proposed text-to-pixel matching operation as follows:

$$\lambda = \mathrm{Match}(t_p, f_p) = \frac{t_p \cdot f_p}{\|t_p\|_2 \times \|f_p\|_2}, \tag{4}$$

where t_p is projected text embeddings with a shape of $K \times D$ (K is the number of categories), f_p is projected feature maps with a shape of $D \times H \times W$, and λ is the text-to-pixel matched embeddings with a shape of $K \times H \times W$.

3.3 The WeakCLIP Framework

The proposed WeakCLIP framework, as shown in Fig. 2, consists of the learnable prompt, pyramid adapter, co-attention matching module, text-guided decoder, and WSSS losses. First, we introduce the learnable prompt we used to extract WSSS-specific text descriptions. Next, we utilize the pyramid adapter to extract fine-grained visual features for WSSS. Then, we perform co-attention matching to model the text-to-pixel relationships as text guidance. Following this, we use a text-guided decoder to incorporate text guidance with adapter output features. Last, we introduce the WSSS losses.

3.3.1 Learnable Prompt for Text Representation

Prompt engineering is important for vision-language models, i.e., CLIP. CLIP-ES (Lin et al., 2022) and CLIMS (Xie

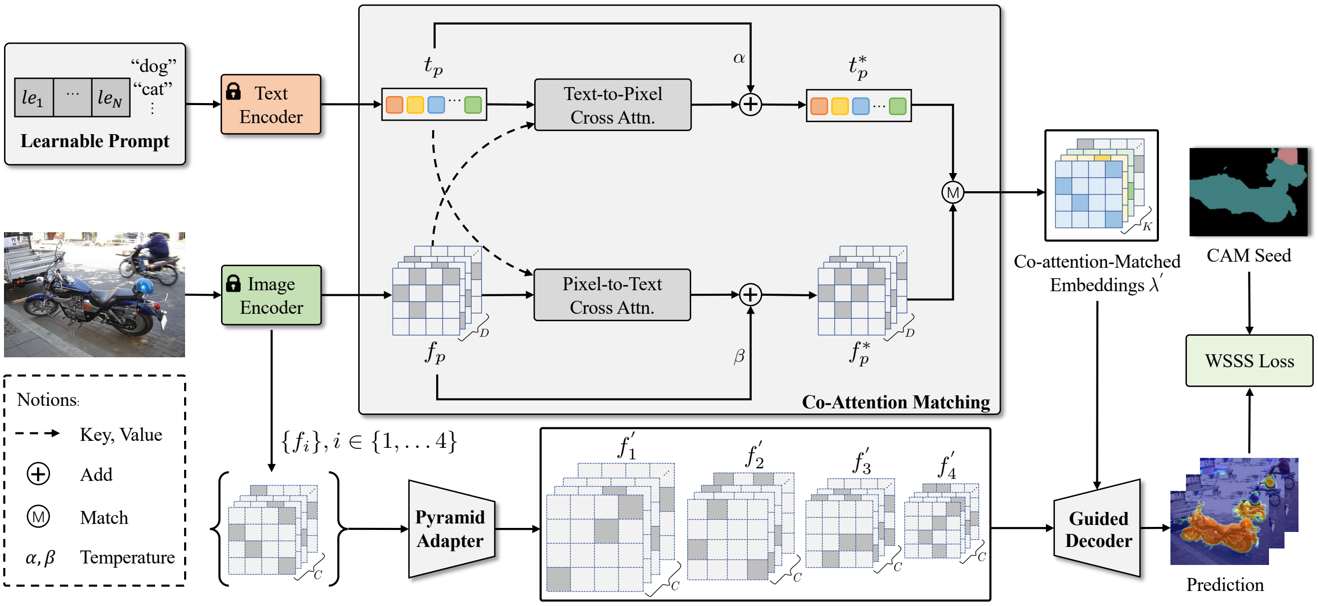


Fig. 2 The training scheme for the proposed WeakCLIP. Firstly, we use the learnable embeddings $\{le_n\}(n \in 1, \ldots, N)$ to enhance the class text input. Next, text and image inputs are translated into two kinds of projected embeddings with the same dimension by CLIP encoders. Then, we apply co-attention matching on the cross-domain embeddings to get co-attention-matched embeddings as text guidance. Subsequently, we use the pyramid adapter to learn WSSS-specific fine-grained image representations and decode image representations with the help of text guidance. Finally, we supervise WeakCLIP with coarse CAM seeds.

et al., 2022) use fixed prompt templates to construct text input for a whole image. However, finding a proper fixed template that could satisfy the need for a pixel-level understanding of WSSS is difficult. Inspired by CoOp (Zhou et al., 2021) and CLIP-Adapter (Gao et al., 2021), we propose learnable embeddings as adaptive prompts to address this problem. We first tokenize and embed the class text into K class text embeddings $\{\mathrm{CLASS}_k\}(k \in 1, \ldots, K)$, each of which has a shape of $L \times C$. Here, L is the context length of class text tokens, and C is the transformer width. Next, we randomly initialize N learnable embeddings $\{le_n\}(n \in 1, \ldots, N)$ as the learnable prompts, each of which has a shape of $1 \times C$. Subsequently, the learnable embeddings are spliced in front of the class text embeddings and they are used as the input embeddings $\{t_k\}(k \in 1, \ldots, K)$, each of which has a shape of $(N + L) \times C$. Specifically, the input embeddings $\{t_k\}(k \in 1, \ldots, K)$ given to the text encoder are designed with the following form:

$$t_k = [le_1, le_2, \ldots, le_N, \mathrm{CLASS}_k]. \tag{5}$$

Then, the text encoder processes the input embeddings $\{t_k\}(k \in 1, \ldots, K)$ and selects the `[EOT]` tokens to represent the corresponding class. Finally, we splice the selected `[EOT]` tokens into a whole and project them to t_p, which has a shape of $K \times D$.

$$t_p = \mathrm{Proj}(\Phi_{\text{text}}(\{t_k\})), \tag{6}$$

where Φ_{text} represents the CLIP text encoder.

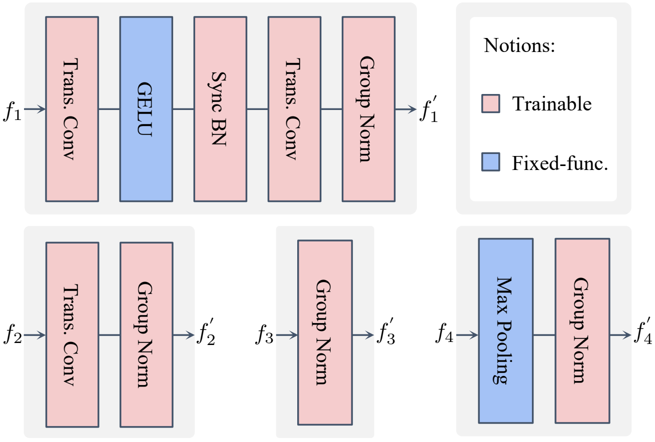


Fig. 3 The structure of our proposed pyramid adapter. The resolution of input features $\{f_i\}(i \in \{1, \ldots, 4\})$ is $\frac{1}{16}$ of the input image. Through the independent processing of each stage, the proposed pyramid adapter outputs features $\{f_i'\}(i \in \{1, \ldots, 4\})$ with resolutions of $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$.

3.3.2 Pyramid Adapter for Image Representation

Limited by the training objective of CLIP, its visual encoder concentrates more on the whole image content than the foreground objects defined by the WSSS task. Besides, another problem stems from the low-resolution problem of CLIP ViT, which also limits its performance in fine-grained understanding. To tackle these problems, we propose a pyramid adapter that operates independently of the CLIP image encoder, facilitates the incremental acquisition of knowledge in the WSSS domain, and employs a hierarchical encoding approach to capture multi-level features.

As shown in Fig. 3, the proposed pyramid adapter comprises lightweight parameters, allowing for independent feature processing across different resolutions. The pyramid adapter first takes feature maps $\{f_i\}(i \in \{1, \ldots, 4\})$ from various transformer layers as input. Subsequently, it performs up-sampling on f_1 and f_2 to the resolutions of $\frac{1}{4}$ and $\frac{1}{8}$ of the original image, respectively, utilizing transposed convolutional operators (Zeiler et al., 2010). Then, the feature map f_4 is down-sampled to $\frac{1}{32}$ resolution of the original image using max pooling operators. Finally, the pyramid adapter generates a set of features $\{f_i^{'}\}(i \in \{1, \ldots, 4\})$ across different resolutions, effectively incorporating both low-level details and high-level representations. Notably, by training the lightweight adapter for pyramid visual representation learning, we avoid fine-tuning CLIP which may destroy the pre-trained knowledge.

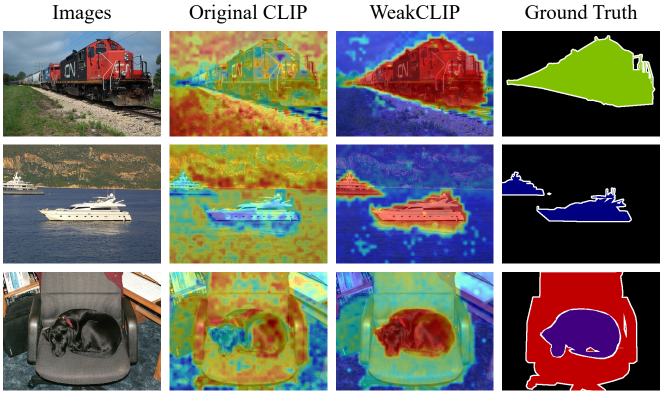


Fig. 4 Comparison of matching results. Original CLIP has many noisy matching activations in the background regions. The proposed WeakCLIP with co-attention matching module provides the text-to-pixel matching for WSSS and achieves better activation results.

3.3.3 Co-Attention Matching for Text Guidance

Text-to-pixel matching is the key motivation of the proposed WeakCLIP to make full use of CLIP pre-trained knowledge. As shown in Fig. 4, since the original CLIP only has the modeling of text-to-image matching, directly using text-to-pixel matching leads to noisy matching results. To solve this problem, we propose co-attention matching to model the text-to-pixel matching for WSSS.

The proposed co-attention matching module utilizes two cross-attention modules to model text-to-pixel relationships and pixel-to-text relationships, respectively. At first, as shown in Fig. 2, the projected text embeddings t_p and projected image embeddings f_p serve as key and value in two different cross-attention modules, respectively. Then, the text and image output from the cross-attention modules are fused with the original t_p and f_p through residual connections.

$$t_p^* = \mathrm{CrossAttn}(t_p, f_p) * \alpha + t_p, \tag{7}$$

$$f_p^* = \mathrm{CrossAttn}(f_p, t_p) * \beta + f_p, \tag{8}$$

where t_p^* and f_p^* represent the updated text embeddings and image embeddings, respectively. $\mathrm{CrossAttn}(\cdot, \cdot)$ denotes the cross-attention operation, while α and β are learnable temperatures that balance the influence of cross-attention output. Lastly, we perform text-to-image matching between the updated text embeddings t_p^* and updated image embeddings f_p^* to get the co-attention-matched embeddings $\lambda^{'}$ as follows:

$$\lambda^{'} = \mathrm{Match}(t_p^*, f_p^*), \tag{9}$$

where the shape of $\lambda^{'}$ is $K \times H \times W$.

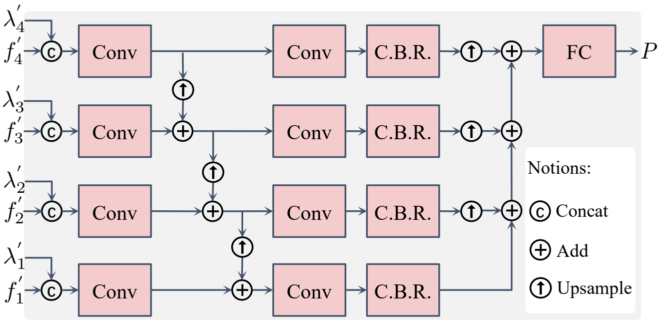


Fig. 5 The structure of the text-guided decoder. We first interpolate the matched embeddings $\lambda^{'}$ to $\{\lambda_i^{'}\}(i \in \{1, \ldots, 4\})$ corresponding to the sizes of the pyramid image features $\{f_i^{'}\}(i \in \{1, \ldots, 4\})$. Then we decode $\{\lambda_i^{'}\}$ and $\{f_i^{'}\}$ hierarchically to get the final prediction P. Notably, the "C.B.R." block represents the "Conv-BN-ReLU" layer.

3.3.4 Text-Guided Decoder for Hierarchical Decoding

To further address the resolution limitation associated with CLIP ViT-B, and integrate adapter output features $\{f_i^{'}\}$ and co-attention-matched embeddings $\lambda^{'}$, we introduce a text-guided decoder as shown in Fig. 5.

In the text-guided decoder $\mathrm{Decoder_g}$, we first interpolate the co-attention-matched embeddings $\lambda^{'}$ to sizes corresponding to the adapter output features $\{f_i^{'}\}$. $\{\lambda_i^{'}\}$ denotes the collections of co-attention-matched embeddings at different spatial sizes. We then concatenate the corresponding image features from $\{f_i^{'}\}$ with the co-attention-matched embeddings from $\{\lambda_i^{'}\}$, and finally decode them to obtain the segmentation prediction P.

$$P = \text{Decoder}_{\text{g}}\left(\{f_i^{'}\}, \{\lambda_i^{'}\}\right), i \in \{1, \ldots 4\}. \tag{10}$$

The hierarchical fusion of adapter output features and co-attention-matched embeddings in WeakCLIP results in more robust segmentation predictions, particularly in segmenting at the details. During training, we compute the WSSS losses using P and CAM seeds. During inference, we apply the $\arg\max$ operation to P to obtain the final segmentation result S as follows:

$$S = \arg\max(P). \tag{11}$$

3.3.5 WSSS Losses

For all our experiments, we adopt the WSSS losses used in DSRG (Huang et al., 2018). The WSSS losses comprise two components: a balanced seeding loss and a boundary loss. The balanced seeding loss computes the weighted cross-entropy loss between the segmentation prediction P and the CAM seeds. The boundary loss first applies conditional random field (CRF) (Krähenbühl and Koltun, 2011) processing to refine the object boundaries in the segmentation prediction P. It then calculates the Kullback-Leibler divergence loss between the CRF-refined results and the segmentation prediction P.

In WSSS tasks, the balanced seed loss function is commonly employed to quantify the discrepancy between predictions and ground truth. It utilizes two normalization coefficients to balance the loss contribution from the foreground and background. Let K represent the collection of classes in the image (excluding the background), and $\bar{K}$ denote the background class. We define M_k as the set of pixels classified as class k, and $P_{u,k}$ as the segmentation prediction for class k at position u. The balanced seeding loss is defined as follows:

$$\begin{aligned} L_{\text{seed}} &= -\frac{1}{\sum_{k\in K}|M_k|}\sum_{k\in K}\sum_{u\in M_k}\log P_{u,k} \\ &\quad -\frac{1}{\sum_{k\in \bar{K}}|M_k|}\sum_{k\in \bar{K}}\sum_{u\in M_k}P_{u,k}. \end{aligned} \tag{12}$$

The Kullback-Leibler divergence is a measure of dissimilarity between two probability distributions. In our case, we utilize the Kullback-Leibler divergence to quantify the disparity between the segmentation prediction and the outcome of CRF refinement. Given the original image I and the segmentation prediction P as inputs, we denote the CRF-processed result as $\Psi_{u,k}(I,P)$. Here, n represents the total number of positions. The boundary loss function is defined as follows:

$$L_{\text{boundary}} = \frac{1}{n}\sum_{u=1}^{n}\sum_{k\in K}\Psi_{u,k}(I,P)\cdot\log\frac{\Psi_{u,k}(I,P)}{P_{u,k}}. \tag{13}$$

3.4 Pseudo Mask Generation and Retrain

We generate high-quality pseudo masks using the trained WeakCLIP network and subsequently perform retraining. To extract the pseudo masks from the segmentation results produced by WeakCLIP, we employ a straightforward approach. When the inference result S includes a category that is not present in the image-level label, we assign an unknown label of 255 to it.

$$\text{PseudoMask}(S) = \begin{cases} S_{h,w}, & S_{h,w} \in K \\ 255, & S_{h,w} \notin K \end{cases}, \tag{14}$$

where h and w represent the image coordinates, S is the inference result obtained from WeakCLIP, and K is the set containing the class labels corresponding to the images. This approach helps prevent incorrect labels from misleading the retraining process.

Finally, we utilize the high-quality pseudo masks generated by WeakCLIP to perform fully-supervised segmentation. To ensure a fair comparison, we follow the retraining settings of MCTformer (Xu et al., 2022) and employ the DeepLabv1 (Wu et al., 2019) network architecture.

4 Experiments

4.1 Datasets and Baseline

Datasets: We evaluate our approach on the PASCAL VOC 2012 dataset (Everingham et al., 2010) and the COCO 2014 dataset (Lin et al., 2014). The PASCAL VOC 2012 dataset consists of 20 foreground objects and one background object. It is divided into three parts: a training set with 1464 images, a validation set with 1449 images, and a test set with 1456 images. To increase the training set, we incorporate the SBD annotations (Hariharan et al., 2011), resulting in a total of 10582 training images. The COCO14 dataset contains 90 categories, with 80 valid foreground objects for segmentation. It has a validation set with 40137 images and a train set with 82081 images. We use the mean intersection over union (mIoU) metric to evaluate the segmentation performance.

Baseline: We use the CAM seeds generated by MCTformer (Xu et al., 2022) as a baseline. We train WeakCLIP model with MCTformer seeds, and generate high-quality pseudo masks.

4.2 Implementation Details

In our experiments, we first refine the MCTformer CAM seeds by the proposed WeakCLIP method. Specifically, we employ ViT-B (Dosovitskiy et al., 2020) as the backbone architecture and use the AdamW (Loshchilov and Hutter, 2017) optimizer. We use a learning rate of 1e-4 and weight decay of 3e-5. During inference, we apply multi-scale and flip strategies, as well as dense CRF for post-processing.

After obtaining the pseudo masks, we proceed with the retraining step using the DeeplabV1 (Chen et al., 2014) framework based on the ResNet38 (Wu et al., 2019).

4.3 Comparison with State-of-the-arts

4.3.1 PASCAL VOC 2012

We present the quantitative results of CAM and pseudo masks for PASCAL VOC 2012 in Table 1. In the second column, it can be observed that the CAM supervision of WeakCLIP is the same as MCTformer (Xu et al., 2022), but lower than ViT-PCM by 1.9% mIoU. The third column demonstrates the quality of the pseudo masks obtained through CAM refinement. Our results exhibit a significant improvement of 8.1% over the baseline MCTformer (Xu et al., 2022) and 5.0% over AMN (Lee et al., 2022b).

Table 1 Evaluation of the CAM and the corresponding pseudo segmentation ground-truth mask (Mask) in terms of mIoU (%) on the PASCAL VOC 2012 $train$ set. We mark the best results in bold.

Method	CAM	Mask
BES(Chen et al., 2020a)	49.6	67.2
SC-CAM(Chang et al., 2020)	50.9	63.4
SEAM(Wang et al., 2020)	55.4	63.6
CDA(Su et al., 2021)	58.4	66.4
CONTA(Zhang et al., 2020b)	56.2	67.9
AdvCAM(Lee et al., 2021b)	55.6	69.9
ECS-Net(Sun et al., 2021)	56.6	67.8
OC-CSE(Kweon et al., 2021)	56.0	66.9
CPN(Zhang et al., 2021a)	57.4	67.8
RIB(Lee et al., 2021a)	56.5	70.6
AMR(Qin et al., 2022)	56.8	69.7
VWE(Ru et al., 2022)	57.3	71.4
CLIMS(Xie et al., 2022)	56.6	70.5
SIPE(Chen et al., 2022a)	58.6	-
AdvCAM + W-OoD(Lee et al., 2022a)	59.1	72.1
AMN(Lee et al., 2022b)	62.1	72.2
ViT-PCM(Rossetti et al., 2022)	**63.6**	67.1
AEFT(Yoon et al., 2022)	56.0	71.0
ACR(Kweon et al., 2023)	65.5	70.9
Baseline and our WeakCLIP.		
MCTformer(Xu et al., 2022)	61.7	69.1
WeakCLIP	61.7	**77.2**+8.1

Table 2 Evaluation of the final segmentation results in terms of mIoU (%) on the PASCAL VOC 2012 val and $test$ sets. The $Sup.$ column denotes the type of supervision used for training including full supervision ($\mathcal{F}$), image-level labels ($\mathcal{I}$), saliency maps ($\mathcal{S}$), and bounding box labels ($\mathcal{B}$). The † indicates the use of the improved ViT pre-trained model. We mark the best WSSS results in bold.

Method	Backbone	$Sup.$	val	$test$
Fully-supervised semantic segmentation (FSSS) methods.				
DeepLabV2(Chen et al., 2017)	ResNet101	$\mathcal{F}$	77.7	79.7
WR38(Wu et al., 2019)	ResNet38		80.8	82.5
WSSS methods with bounding box.				
BCM(Song et al., 2019)	ResNet101	$\mathcal{I}+\mathcal{B}$	70.2	-
BBAM(Lee et al., 2021c)	ResNet101		73.7	73.7
WSSS methods with saliency map.				
ICD(Fan et al., 2020)	ResNet101	$\mathcal{I}+\mathcal{S}$	67.8	68.0
EPS(Lee et al., 2021d)	ResNet101		71.0	71.8
L2G(Jiang et al., 2022)	ResNet101		72.1	71.7
WSSS methods with only image-level labels.				
BES(Chen et al., 2020a)	ResNet101	$\mathcal{I}$	65.7	66.6
SC-CAM(Chang et al., 2020)	ResNet101		66.1	65.9
SEAM(Wang et al., 2020)	ResNet38		64.5	65.7
CDA(Su et al., 2021)	ResNet38		66.1	66.8
CONTA(Zhang et al., 2020b)	ResNet38		66.1	66.7
AdvCAM(Lee et al., 2021b)	ResNet101		68.1	68.0
ECS-Net(Sun et al., 2021)	ResNet38		66.6	67.6
PMM(Li et al., 2021b)	Res2Net101		70.0	70.5
OC-CSE(Kweon et al., 2021)	ResNet38		68.4	68.2
ReCAM(Chen et al., 2022b)	ResNet101		68.5	68.4
CPN(Zhang et al., 2021a)	ResNet38		67.8	68.5
RIB(Lee et al., 2021a)	ResNet101		68.3	68.6
AMR(Qin et al., 2022)	ResNet101		68.8	69.1
VWE(Ru et al., 2022)	ResNet101		70.6	70.7
URN(Li et al., 2021a)	Res2Net101		71.2	71.5
CLIMS(Xie et al., 2022)	ResNet101		70.4	70.0
SANCE(Li et al., 2022)	ResNet101		70.9	72.2
SIPE(Chen et al., 2022a)	ResNet101		68.8	69.7
W-OoD(Lee et al., 2022a)	ResNet101		70.7	70.1
AMN(Lee et al., 2022b)	ResNet101		69.5	69.6
ViT-PCM(Rossetti et al., 2022)	ResNet101		70.3	70.9
AEFT(Yoon et al., 2022)	ResNet38		70.9	71.7
ToCo(Ru et al., 2023)	ViT-B		69.8	70.5
OCR(Cheng et al., 2023)	ResNet38		72.7	72.0
ACR(Kweon et al., 2023)	ResNet38		72.4	72.4
Baseline and our WeakCLIP.				
MCTformer(Xu et al., 2022)	ResNet38	$\mathcal{I}$	71.9	71.6
WeakCLIP	ResNet38		**74.0**+2.1	**73.8**+2.2

To further validate the quality of the pseudo masks generated by WeakCLIP, we fully train a segmentation network, i.e., DeepLabV1 (Chen et al., 2014), on the PASCAL VOC 2012 dataset using the refined pseudo masks. As shown in Table 2, WeakCLIP achieves 74.0% and 73.8% mIoU on the PASCAL VOC 2012 val and $test$ sets, respectively. Specifically, compared to our baseline method, MCTformer (Xu et al., 2022), WeakCLIP outperforms it by 2.1% and 2.2% mIoU on the val and $test$ sets, respectively. Moreover, compared to the other methods with only image-level supervision, WeakCLIP exhibits superior performance. It attains a 1.3% higher mIoU than OCR (Cheng et al., 2023) on the val set and a 1.4% higher mIoU than ACR (Kweon et al., 2023) on the $test$ set. In addition, compared to methods with additional saliency map supervision or box supervision, e.g., L2G (Jiang et al., 2022) and BBAM (Lee et al., 2021c), our method also achieves superior performance. The results demonstrate that WeakCLIP can effectively improve the quality of pseudo masks and achieve state-of-the-art results on the PASCAL VOC 2012 dataset.

Table 3 Evaluation of the final segmentation results in terms of mIoU (%) on the COCO 2014 *val* set. We mark the best WSSS results in bold.

Method	Backbone	*Sup.*	*val*
WSSS methods with saliency map.			
EPS(Lee et al., 2021d)	ResNet101	$\mathcal{I}+\mathcal{S}$	35.7
AuxSegNet(Xu et al., 2021b)	ResNet38		33.9
WSSS methods with only image-level labels.			
SEAM(Wang et al., 2020)	ResNet38	$\mathcal{I}$	31.9
CDA(Su et al., 2021)	ResNet38		33.2
CONTA(Zhang et al., 2020b)	ResNet38		32.8
PMM(Li et al., 2021b)	ScaleNet101		40.2
OC-CSE(Kweon et al., 2021)	ResNet38		36.4
RIB(Lee et al., 2021a)	ResNet101		43.8
VWE(Ru et al., 2022)	ResNet101		36.2
URN(Li et al., 2021a)	Res2Net101		41.5
SANCE(Li et al., 2022)	ResNet101		44.7
SIPE(Chen et al., 2022a)	ResNet38		43.6
AMN(Lee et al., 2022b)	ResNet101		44.7
ViT-PCM(Rossetti et al., 2022)	ResNet101		45.0
AEFT(Yoon et al., 2022)	ResNet38		44.8
ToCo(Ru et al., 2023)	ViT-B		41.3
OCR(Cheng et al., 2023)	ResNet38		42.5
ACR(Kweon et al., 2023)	ResNet38		45.3
Baseline and our WeakCLIP.			
MCTformer(Xu et al., 2022)	ResNet38	$\mathcal{I}$	42.0
WeakCLIP	ResNet38		**46.1**+4.1

4.3.2 COCO 2014

As shown in Table 3, WeakCLIP achieves 46.1% mIoU on the most challenging benchmark, COCO 2014 *val* set. Specifically, compared to our baseline method, MCTformer (Xu et al., 2022), WeakCLIP outperforms it by 4.1% mIoU on the *val* set. Moreover, compared to the other methods with only image-level supervision, WeakCLIP exhibits superior performance. It attains a 0.8% higher mIoU than ACR (Kweon et al., 2023) on the *val* set. The results also show that WeakCLIP achieves state-of-the-art results on the COCO 2014 dataset.

4.4 Ablation Studies

4.4.1 Baselines for Ablation Studies

As shown in Fig. 4, we present three baselines as basis. The first baseline is MCTformer, which achieves an mIoU of 60.0% on the PASCAL VOC 2012 *val* set. For the baseline of fully fine-tuning CLIP, we train a model with unfixed CLIP encoders and a segmentation head. This baseline achieves a mIoU of only 58.1% on the validation set, indicating that fine-tuning CLIP with coarse CAM disrupts its powerful representation. In another baseline, we fix the weights of CLIP encoders and directly utilize the text-to-pixel matching results as the pixel-level semantics. However, this approach yields a poor result, with an mIoU of 12.3% on the validation set, as the pre-training objective of CLIP lacks the modeling of text-to-pixel matching.

4.4.2 Improvements of WeakCLIP Components

To analyze the improvements brought by our proposed WeakCLIP components, we present the quantitative results of these components on the VOC 2012 *val* set in Table 4.

Table 4 Ablation study for WeakCLIP components on PASCAL VOC 2012 *val* set. We mark the best WSSS results in bold.

CLIP Encoders	Co-Attn. Matching	Learnable Embed.	Pyramid Adapter	Text-Guided Decoder	mIoU
(Baseline: CAM from MCTformer.)					60.0
Direct apply CLIP to WSSS.					
Unfixed					58.1
Fixed					12.3
Our WeakCLIP.					
Fixed	✓				67.4
Fixed	✓	✓			68.9
Fixed	✓	✓	✓		70.3
Fixed	✓	✓	✓	✓	**72.6**

First, the co-attention matching module models the relationships between text and pixels and improves the results to 67.4% mIoU on the *val* set. Next, by employing the learnable prompt to capture WSSS-specific text descriptions, the text-to-pixel representations are significantly improved, resulting in an mIoU of 68.9% on the *val* set. Then, by incorporating the pyramid adapter to learn WSSS-specific multi-level image representations, we observe an increase in performance to 70.3% mIoU on the *val* set. Finally, the text-guided decoder, which integrates detailed information and text-to-pixel guidance in a hierarchical manner, enhances the performance to 72.6% mIoU on the *val* set.

Table 5 Ablation study for the number of learnable embedding on PASCAL VOC 2012 *val* set. We mark the best WSSS results in bold.

Number of Learnable Embeddings	2	4	8	16
mIoU	72.2	72.3	**72.6**	72.3

Table 6 Comparison of per-class segmentation results in terms of IoUs on the PASCAL VOC 2012 val set.

Method	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow
MCTformer $_{\text{CVPR22}}$(Xu et al., 2022)	91.9	78.3	39.5	89.9	55.9	76.7	81.8	**79.0**	90.7	32.6	**87.1**
WeakCLIP (Ours)	**92.7**	**87.1**	**40.6**	**89.9**	**63.0**	**78.3**	**86.8**	77.9	**90.7**	**33.0**	84.8
Method	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	**mIoU**
MCTformer $_{\text{CVPR22}}$(Xu et al., 2022)	**57.2**	87.0	**84.6**	77.4	79.2	55.1	**89.2**	47.2	70.4	58.8	71.9
WeakCLIP (Ours)	48.4	**88.2**	83.8	**78.4**	**81.4**	**64.9**	87.8	**53.6**	**76.4**	**66.6**	**74.0**

Table 7 Comparison of per-class segmentation results in terms of IoUs on the PASCAL VOC 2012 $test$ set.

Method	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow
MCTformer $_{\text{CVPR22}}$(Xu et al., 2022)	92.3	84.4	37.2	82.8	**60.0**	**72.8**	78.0	79.0	89.4	31.7	**84.5**
WeakCLIP (Ours)	**92.9**	**88.4**	**40.6**	**88.3**	57.6	71.8	**82.6**	**80.0**	**89.9**	**33.1**	82.6
Method	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	**mIoU**
MCTformer $_{\text{CVPR22}}$(Xu et al., 2022)	**59.1**	85.3	**83.8**	79.2	**81.0**	53.9	85.3	60.5	65.7	57.7	71.6
WeakCLIP (Ours)	57.6	**86.0**	82.9	**83.9**	79.6	**66.6**	**86.8**	**60.9**	**72.8**	**64.2**	**73.8**

4.4.3 Number of Learnable Embeddings

To examine the impact of the number of learnable embeddings, we perform ablation studies and present the results in Table 5. From the table, it is evident that the best performance is achieved when the number of learnable embeddings is set to 8, while slightly inferior results are obtained with other numbers. These findings indicate that the number of learnable embeddings does indeed influence the results, and selecting an appropriate number of learnable embeddings can enhance the text representation in weakly-supervised semantic segmentation scenarios.

4.5 Visualization

4.5.1 Initial value of learnable temperature

We also conduct ablation studies to explore the impact of the initial value of the learnable temperature in co-attention matching. The results are summarized in Table 8. We can observe that using an initial value of 1e-1 for the learnable temperature leads to the best performance.

Table 8 Ablation study for the initial value of learnable temperatures in co-attention matching on PASCAL VOC 2012 val set.

Initial Value of Temperatures	1	1e-1	1e-2	1e-3
mIoU	72.5	**72.6**	72.3	72.2

4.6 Per-class Semantic Segmentation Results

In Tables 6 and 7, we present the per-class segmentation results on the val and $test$ sets of PASCAL VOC 2012. Additionally, Table 9 shows the per-class segmentation results on the validation set of COCO 2014. We compare the performance of our proposed WeakCLIP with baseline method, MCTformer (Xu et al., 2022). The results indicate that WeakCLIP achieves superior performance in most categories, demonstrating its effectiveness in the weakly-supervised semantic segmentation (WSSS) domain.

To evaluate the quality of the segmentation results obtained by our proposed method, we conduct a qualitative analysis by comparing the pseudo masks generated by MCTformer and WeakCLIP. Fig. 6 illustrates the comparison of pseudo masks generated using the CAM of MCTformer for both methods. It is evident that WeakCLIP method generates more accurate and precise semantic information, particularly for object locations that were missed or inaccurately identified by MCTformer.

Furthermore, we provide visualizations of the segmentation results obtained after retraining on the PASCAL VOC 2012 validation set. Fig. 7 showcases the original images, the segmentation results produced by WeakCLIP, and the corresponding ground truth (GT). The visualizations demonstrate that our method achieves accurate segmentation results for both indoor and outdoor scenes.

5 Conclusion

Dealing with the noisy and sparse class activation map (CAM) seeds in current Weakly-supervised Semantic Seg-

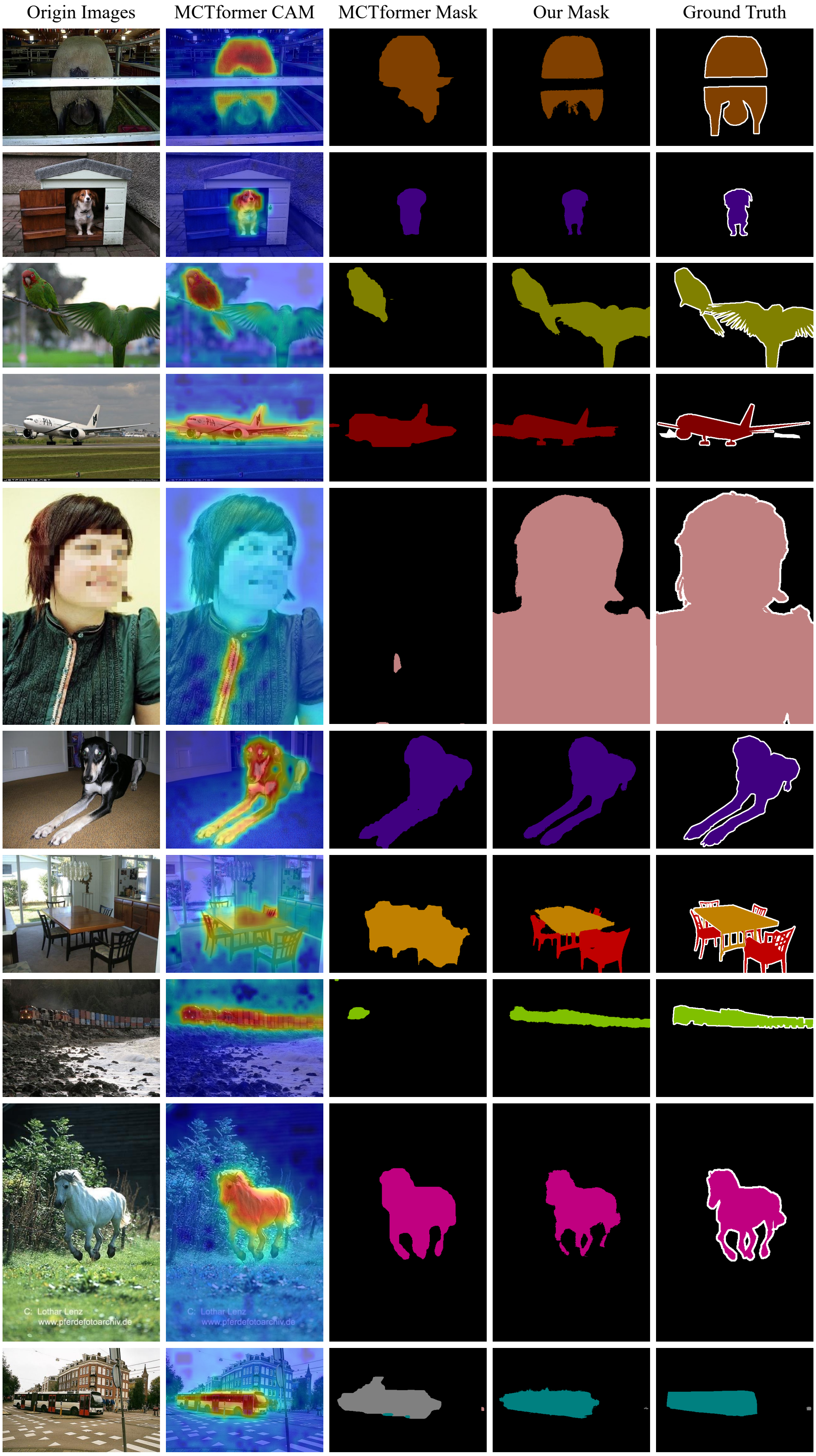


Fig. 6 Comparison of the pseudo mask on the PASCAL VOC 2012 $train$ set.

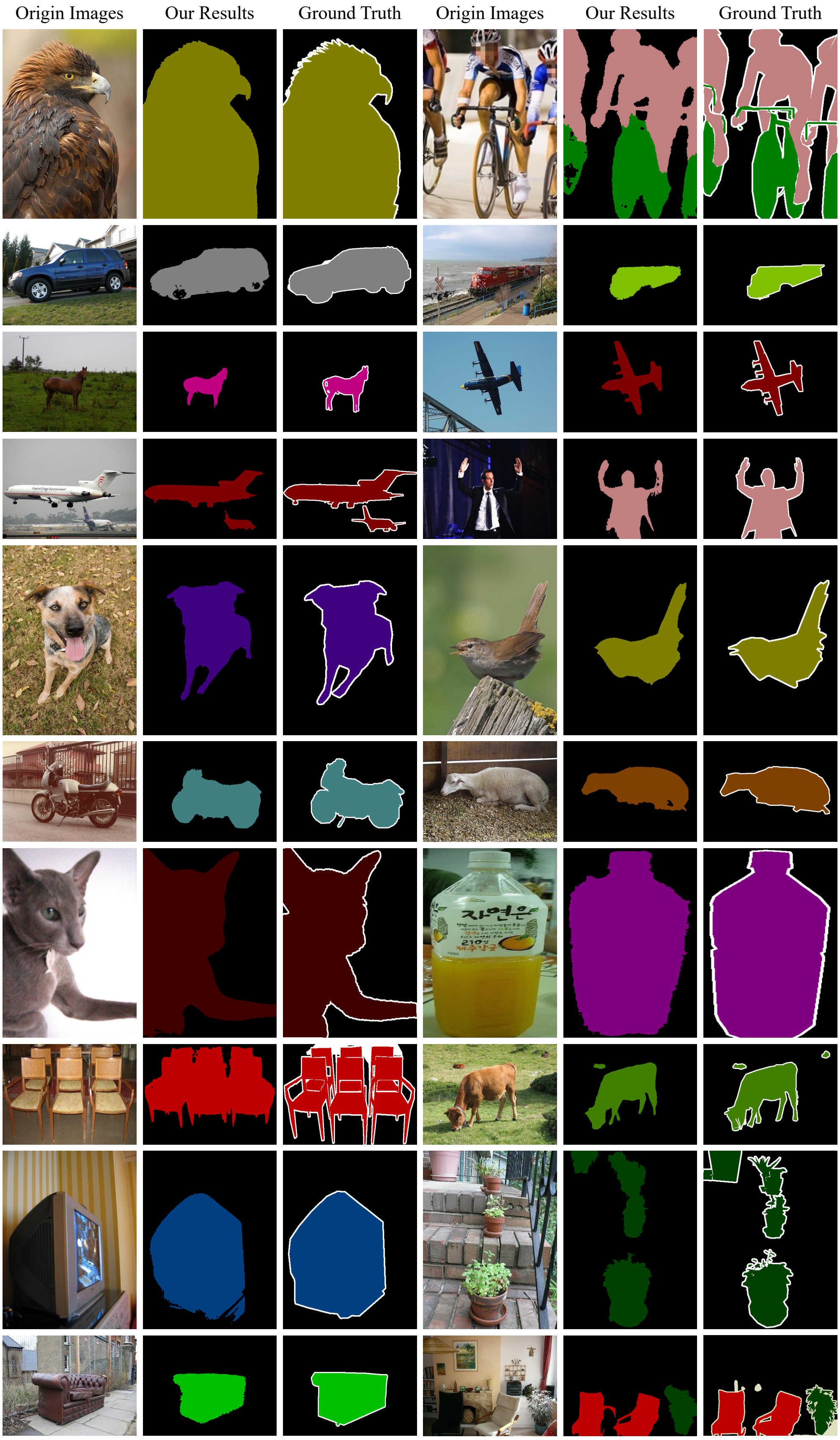


Fig. 7 Segmentation visualization results on the PASCAL VOC 2012 *val* set.

Table 9 Comparison of per-class segmentation results in terms of IoUs on the COCO 2014 *val* set. We mark the best results in bold.

Class	MCTformer CVPR22(Xu et al., 2022)	WeakCLIP (Ours)	Class	MCTformer CVPR22(Xu et al., 2022)	WeakCLIP (Ours)
background	82.4	**82.4**	wine glass	**27.0**	19.3
person	**62.6**	59.8	cup	**29.0**	26.9
bicycle	47.4	**49.7**	fork	**23.4**	15.0
car	47.2	**47.7**	knife	12.0	**18.3**
motorcycle	63.7	**68.7**	spoon	**6.6**	2.6
airplane	**64.7**	55.8	bowl	**22.4**	19.9
bus	64.5	**72.0**	banana	63.2	**71.7**
train	64.5	**69.1**	apple	44.4	**55.5**
truck	44.8	**51.1**	sandwich	39.7	**43.7**
boat	42.3	**45.3**	orange	63.0	**69.1**
traffic light	49.9	**54.9**	broccoli	51.2	**66.7**
fire hydrant	73.2	**77.8**	carrot	40.0	**48.9**
stop sign	76.6	**78.2**	hot dog	53.0	**53.6**
parking meter	64.4	**71.2**	pizza	62.2	**65.1**
bench	32.8	**43.1**	donut	55.7	**67.5**
bird	62.6	**67.1**	cake	47.9	**58.0**
cat	78.2	**81.3**	chair	22.8	**23.3**
dog	68.2	**72.5**	couch	35.0	**39.1**
horse	65.8	**69.4**	potted plant	13.5	**15.9**
sheep	70.1	**75.0**	bed	48.6	**52.1**
cow	68.3	**75.9**	dining table	**12.9**	3.2
elephant	81.6	**83.7**	toilet	63.1	**69.1**
bear	80.1	**82.6**	tv	47.9	**52.5**
zebra	83.0	**83.9**	laptop	49.5	**54.9**
giraffe	76.9	**80.0**	mouse	13.4	**15.0**
backpack	14.6	**20.2**	remote	41.9	**48.1**
umbrella	61.7	**67.1**	keyboard	49.8	**50.5**
handbag	4.5	**9.4**	cellphone	54.1	**59.0**
tie	25.2	**32.8**	microwave	38.0	**47.5**
suitcase	46.8	**53.0**	oven	29.9	**37.5**
frisbee	43.8	**62.9**	toaster	0.0	0.0
skis	12.8	**13.5**	sink	28.0	**28.8**
snowboard	31.4	**35.7**	refrigerator	40.1	**52.2**
sports ball	9.2	**23.7**	book	32.2	**32.2**
kite	26.3	**44.1**	clock	**43.2**	36.0
baseball bat	**0.9**	0.5	vase	22.6	**28.1**
baseball glove	0.7	**3.4**	scissors	32.9	**42.0**
skateboard	7.8	**11.7**	teddy bear	61.9	**66.5**
surfboard	46.5	**49.0**	hair drier	0.0	0.0
tennis racket	1.4	**3.6**	toothbrush	12.2	**20.4**
bottle	**31.1**	25.7	**mIoU**	42.0	**46.1**

mentation (WSSS) methods is a significant challenge. In this regard, we propose a novel scheme called WeakCLIP that leverages the knowledge from pre-trained CLIP models to enhance the CAM refinement process of WSSS networks. Our proposed WeakCLIP framework adopts a novel text-to-pixel matching paradigm and effectively tackles three key problems associated with integrating CLIP into WSSS. Experimental results on the widely-used PASCAL VOC 2012 and COCO 2014 datasets demonstrate the significant improvements achieved by WeakCLIP compared to previous WSSS methods. The introduction of the WeakCLIP paradigm, which harnesses large-scale vision language pre-training, holds promise for advancing solutions to the WSSS problem. In our future work, we will explore the more advanced large-scale CLIP in boosting the pixel-level understanding of WSSS.

References

Ahn J, Kwak S (2018) Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4981–4990

Ahn J, Cho S, Kwak S (2019) Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2209–2218

Bearman A, Russakovsky O, Ferrari V, Fei-Fei L (2016) What's the point: Semantic segmentation with point supervision. In: European conference on computer vision, Springer, pp 549–565

Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, Joulin A (2021) Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 9650–9660

Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6299–6308

Chang YT, Wang Q, Hung WC, Piramuthu R, Tsai YH, Yang MH (2020) Weakly-supervised semantic segmentation via sub-category exploration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8991–9000

Chen L, Wu W, Fu C, Han X, Zhang Y (2020a) Weakly supervised semantic segmentation with boundary exploration. In: European Conference on Computer Vision, Springer, pp 347–362

Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2014) Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:14127062

Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence 40(4):834–848

Chen Q, Yang L, Lai JH, Xie X (2022a) Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4288–4298

Chen T, Kornblith S, Norouzi M, Hinton G (2020b) A simple framework for contrastive learning of visual representations. In: International conference on machine learning, PMLR, pp 1597–1607

Chen Z, Wang T, Wu X, Hua XS, Zhang H, Sun Q (2022b) Class re-activation maps for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 969–978

Cheng Z, Qiao P, Li K, Li S, Wei P, Ji X, Yuan L, Liu C, Chen J (2023) Out-of-candidate rectification for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 23673–23684

Dai J, He K, Sun J (2015) Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 1635–1643

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929

Du Y, Fu Z, Liu Q, Wang Y (2022) Weakly supervised semantic segmentation by pixel-to-prototype contrast. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4320–4329

Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. International journal of computer vision 88(2):303–338

Fan J, Zhang Z, Song C, Tan T (2020) Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4283–4292

Feng J, Wang X, Liu W (2021) Deep graph cut network for weakly-supervised semantic segmentation. Science China Information Sciences 64(3):1–12

Gao P, Geng S, Zhang R, Ma T, Fang R, Zhang Y, Li H, Qiao Y (2021) Clip-adapter: Better vision-language models with feature adapters. arXiv preprint arXiv:211004544

Hariharan B, Arbeláez P, Bourdev L, Maji S, Malik J (2011) Semantic contours from inverse detectors. In: 2011 international conference on computer vision, IEEE, pp 991–998

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9729–9738

Huang Z, Wang X, Wang J, Liu W, Wang J (2018) Weakly-supervised semantic segmentation network with deep seeded region growing. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7014–7023

Jiang PT, Hou Q, Cao Y, Cheng MM, Wei Y, Xiong HK (2019) Integral object mining via online attention accumulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 2070–2079

Jiang PT, Yang Y, Hou Q, Wei Y (2022) L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 16886–16896

Khoreva A, Benenson R, Hosang J, Hein M, Schiele B (2017) Simple does it: Weakly supervised instance and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 876–885

Kolesnikov A, Lampert CH (2016) Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: European conference on computer vision, Springer, pp 695–711

Krähenbühl P, Koltun V (2011) Efficient inference in fully connected crfs with gaussian edge potentials. Advances in neural information processing systems 24

Kweon H, Yoon SH, Kim H, Park D, Yoon KJ (2021) Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 6994–7003

Kweon H, Yoon SH, Yoon KJ (2023) Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 11329–11339

Lee J, Kim E, Lee S, Lee J, Yoon S (2019a) Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5267–5276

Lee J, Kim E, Lee S, Lee J, Yoon S (2019b) Frame-to-frame aggregation of active regions in web videos for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 6808–6818

Lee J, Choi J, Mok J, Yoon S (2021a) Reducing information bottleneck for weakly supervised semantic segmentation. Advances in Neural Information Processing Systems 34:27408–27421

Lee J, Kim E, Yoon S (2021b) Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4071–4080

Lee J, Yi J, Shin C, Yoon S (2021c) Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2643–2652

Lee J, Oh SJ, Yun S, Choe J, Kim E, Yoon S (2022a) Weakly supervised semantic segmentation using out-of-distribution data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 16897–16906

Lee M, Kim D, Shim H (2022b) Threshold matters in wsss: Manipulating the activation for the robust and accurate segmentation model against thresholds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4330–4339

Lee S, Lee M, Lee J, Shim H (2021d) Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5495–5505

Lei J, Li L, Zhou L, Gan Z, Berg TL, Bansal M, Liu J (2021) Less is more: Clipbert for video-and-language learning via sparse sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7331–7341

Li J, Fan J, Zhang Z (2022) Towards noiseless object contours for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 16856–16865

Li Y, Duan Y, Kuang Z, Chen Y, Zhang W, Li X (2021a) Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. arXiv preprint arXiv:211207431

Li Y, Kuang Z, Liu L, Chen Y, Zhang W (2021b) Pseudo-mask matters in weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 6964–6973

Lin D, Dai J, Jia J, He K, Sun J (2016) Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3159–3167

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision, Springer, pp 740–755

Lin Y, Chen M, Wang W, Wu B, Li K, Lin B, Liu H, He X (2022) Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. arXiv preprint arXiv:221209506

Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. arXiv preprint arXiv:171105101

Lu J, Batra D, Parikh D, Lee S (2019) Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems 32

Pathak D, Krahenbuhl P, Darrell T (2015) Constrained convolutional neural networks for weakly supervised segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 1796–1804

Pinheiro PO, Collobert R (2015) From image-level to pixel-level labeling with convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1713–1721

Qin J, Wu J, Xiao X, Li L, Wang X (2022) Activation modulation and recalibration scheme for weakly supervised semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 36, pp 2117–2125

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al. (2021) Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, PMLR, pp 8748–8763

Rao Y, Zhao W, Chen G, Tang Y, Zhu Z, Huang G, Zhou J, Lu J (2021) Denseclip: Language-guided dense prediction with context-aware prompting. arXiv preprint arXiv:211201518

Rossetti S, Zappia D, Sanzari M, Schaerf M, Pirri F (2022) Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX, Springer, pp 446–463

Ru L, Du B, Zhan Y, Wu C (2022) Weakly-supervised semantic segmentation with visual words learning and hybrid pooling. International Journal of Computer Vision 130(4):1127–1144

Ru L, Zheng H, Zhan Y, Du B (2023) Token contrast for weakly-supervised semantic segmentation. arXiv preprint arXiv:230301267

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626

Shen T, Lin G, Liu L, Shen C, Reid I (2017) Weakly supervised semantic segmentation based on co-segmentation. In: BMVC

Song C, Huang Y, Ouyang W, Wang L (2019) Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3136–3145

Su W, Zhu X, Cao Y, Li B, Lu L, Wei F, Dai J (2019) Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:190808530

Su Y, Sun R, Lin G, Wu Q (2021) Context decoupling augmentation for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7004–7014

Sun C, Shrivastava A, Singh S, Gupta A (2017) Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE international conference on computer vision, pp 843–852

Sun K, Shi H, Zhang Z, Huang Y (2021) Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 7283–7292

Tang M, Djelouah A, Perazzi F, Boykov Y, Schroers C (2018) Normalized cut loss for weakly-supervised cnn segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1818–1827

Vernaza P, Chandraker M (2017) Learning random-walk label propagation for weakly-supervised semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7158–7166

Wang Y, Zhang J, Kan M, Shan S, Chen X (2020) Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 12275–12284

Wei Y, Liang X, Chen Y, Shen X, Cheng MM, Feng J, Zhao Y, Yan S (2016) Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE transactions on pattern analysis and machine intelligence 39(11):2314–2320

Wei Y, Feng J, Liang X, Cheng MM, Zhao Y, Yan S (2017) Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1568–1576

Wei Y, Xiao H, Shi H, Jie Z, Feng J, Huang TS (2018) Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7268–7277

Wu Z, Shen C, Van Den Hengel A (2019) Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recognition 90:119–133

Xie J, Hou X, Ye K, Shen L (2022) Clims: cross language image matching for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4483–4492

Xu J, Schwing AG, Urtasun R (2015) Learning to segment under various forms of weak supervision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3781–3790

Xu L, Ouyang W, Bennamoun M, Boussaid F, Sohel F, Xu D (2021a) Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 6984–6993

Xu L, Ouyang W, Bennamoun M, Boussaid F, Sohel F, Xu D (2021b) Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 6984–6993

Xu L, Ouyang W, Bennamoun M, Boussaid F, Xu D (2022) Multi-class token transformer for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4310–4319

Yang J, Sun X, Lai YK, Zheng L, Cheng MM (2018) Recognition from web data: A progressive filtering approach. IEEE Transactions on Image Processing 27(11):5303–5315

Yao Y, Chen T, Xie GS, Zhang C, Shen F, Wu Q, Tang Z, Zhang J (2021) Non-salient region object mining for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2623–2632

Yoon SH, Kweon H, Cho J, Kim S, Yoon KJ (2022) Adversarial erasing framework via triplet with gated pyramid pooling layer for weakly supervised semantic segmentation. In: European Conference on Computer Vision, Springer, pp 326–344

Zeiler MD, Krishnan D, Taylor GW, Fergus R (2010) Deconvolutional networks. In: 2010 IEEE Computer Society Conference on computer vision and pattern recognition, IEEE, pp 2528–2535

Zhang B, Xiao J, Wei Y, Sun M, Huang K (2020a) Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 34, pp 12765–12772

Zhang D, Zhang H, Tang J, Hua XS, Sun Q (2020b) Causal intervention for weakly-supervised semantic segmentation. Advances in Neural Information Processing Systems 33:655–666

Zhang F, Gu C, Zhang C, Dai Y (2021a) Complementary patch for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 7242–7251

Zhang R, Fang R, Gao P, Zhang W, Li K, Dai J, Qiao Y, Li H (2021b) Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:211103930

Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2921–2929

Zhou K, Yang J, Loy CC, Liu Z (2021) Learning to prompt for vision-language models. arXiv preprint arXiv:210901134