Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.059149

ARTICLE

A Weakly Supervised Semantic Segmentation Method Based on Improved Conformer

Xueli Shen and Meng Wang*

School of Software, Liaoning Technical University, Huludao, 125105, China

* Corresponding Author: Meng Wang. Email: wmyyya@126.com

Received: 29 September 2024; Accepted: 12 December 2024; Published: 06 March 2025

ABSTRACT: In the field of Weakly Supervised Semantic Segmentation (WSSS), methods based on image-level annotation face challenges in accurately capturing objects of varying sizes, lacking sensitivity to image details, and having high computational costs. To address these issues, we improve the dual-branch architecture of the Conformer as the fundamental network for generating class activation graphs, proposing a multi-scale efficient weakly-supervised semantic segmentation method based on the improved Conformer. In the Convolution Neural Network (CNN) branch, a cross-scale feature integration convolution module is designed, incorporating multi-receptive field convolution layers to enhance the model's ability to capture long-range dependencies and improve sensitivity to multi-scale objects. In the Vision Transformer (ViT) branch, an efficient multi-head self-attention module is developed, reducing unnecessary computation through spatial compression and feature partitioning, thereby improving overall network efficiency. Finally, a multi-feature coupling module is introduced to complement the features generated by both branches. This design retains the strength of Convolution Neural Network in extracting local details while harnessing the strength of Vision Transformer to capture comprehensive global features. Experimental results show that the mean Intersection over Union of the image segmentation results of the proposed method on the validation and test sets of the PASCAL VOC 2012 datasets are improved by 2.9% and 3.6%, respectively, over the TransCAM algorithm. Besides, the improved model demonstrates a 1.3% increase of the mean Intersections over Union on the COCO 2014 datasets. Additionally, the number of parameters and the floating-point operations are reduced by 16.2% and 12.9%. However, the proposed method still has limitations of poor performance when dealing with complex scenarios. There is a need for further enhancing the performance of this method to address this issue.

KEYWORDS: WSSS; CAM; transformer; CNN; multi-scale feature extraction; lightweight

1 Introduction

In today's era of exploding visual information, we are in constant contact with images. Whether it's the vast number of photos on social media or the monitoring images of harsh weather conditions in reality [1], the understanding and analysis of images are of utmost importance. Semantic segmentation, as a pivotal task in the field of computer vision, is like a magical key that can unlock the abundant semantic information contained within images. In recent years, although traditional semantic segmentation methods based on deep learning have made significant progress [2,3]. However, these methods typically rely on extensive of pixel-level annotated training datasets, the acquisition of which is not only expensive but also time-consuming and labor-intensive. As a result, weakly supervised semantic segmentation (WSSS) technology emerges as the times require. It seeks to train segmentation models through simpler supervision information and has become a major hotspot in current academic research. It only requires some relatively easily obtained

image annotations, such as bounding boxes [4], scribbles [5], point annotations [6], and image-level class labels [7], to achieve efficient semantic segmentation. Since large-scale datasets like ImageNet [8] already provide image-level class labels, this form of supervision is the easiest to access. By integrating WSSS with image-level annotations and other artificial intelligence technologies, it can be applied in fields such as autonomous driving and medical image diagnosis [9], and then promote the development of various fields in the direction of intelligence.

In recent years, research on weakly supervised semantic segmentation has made remarkable progress. Among various methods, a common method is to first utilize a convolutional neural network (CNN) to generate a class activation map (CAM) for locating the position of the target. The class activation map is then employed to produce a pseudo-label for the segmentation network. This pseudo-label is then used as the ground-truth to train the existing highly precise fully supervised image segmentation network, finally obtaining the image segmentation result. However, CAMs generated by CNNs often focus only on activating partial regions of the target object in an image. Although many studies have proposed various methods to generate larger activation regions aligned with the full object region [10–13], the defects when applied to weakly supervised semantic segmentation still have not been directly resolved. Vision Transformer (ViT) [14], which incorporates a multi-head self-attention (MHSA) mechanism inspired by the Transformer model [15], has attained breakthrough performance in image recognition tasks and has subsequently been utilized within the domain of weakly supervised semantic segmentation. Different from convolutional neural networks, Vision Transformer leverages its global attention mechanism to provide a broader global receptive field. When applied within the domain of WSSS, attention maps can be used to capture global features, but there is a defect of lacking local features. Since CNN and ViT-based methods have their own strengths and weaknesses, there has been growing research focused on combining both networks to improve performance by complementing each other. A notable example is TransCAM [16], based on the Conformer [17]. This method uses the class activation map generated by the convolutional neural network and at the same time uses the attention map obtained by Vision Transformer to enhance the obtained class activation map, thereby achieving better results. However, in this method, the convolutional neural network branch uses a traditional feature pyramid structure, where each convolutional block generates feature maps at a single scale, resulting in the loss of small object information and limiting its ability to handle complex scenes with multi-scale objects. In addition, the Vision Transformer branch structure's attention mechanism entails a substantial quadratic computational complexity, resulting in a heavy computational demand problem.

To address these issues, we improve main modules in the dual-branch structure of the Conformer network and introduces a multi-feature coupling module that leverages the distinct strengths of both convolutional neural networks and Vision Transformers to their fullest potential.

The following points outline the primary contributions of this paper:

(1) In the convolutional neural network branch, a multi-feeling field convolutional layer is introduced to construct a cross-scale feature integration convolutional module, which augments the capacity of the CNN branch to grasp long-distance dependencies, making the model perform better in coping with targets of different sizes, capturing more details and improving the sensitivity to multi-scale targets.

(2) An efficient multi-head self-attention module is proposed to replace the self-attention module in the Vision Transformer branch. Spatial compression and feature partitioning operations are introduced into the traditional self-attention module to perform spatial compression on keys and values tensors to reduce unnecessary computations, thereby improving the overall efficiency of the network.

(3) For the feature complementarity of the above two branches, a multi-feature coupling module is constructed to fully utilize the advantages of features from different resolutions and levels, while retaining the

respective strengths of CNN and ViT in local and global feature extraction. Experimental results show that, compared to existing mainstream methods, the proposed approach notably diminishes the computational complexity of WSSS while enhancing segmentation precision.

2 Related Work

2.1 Weakly Supervised Semantic Segmentation

In fully supervised semantic segmentation models based on deep learning, fully convolutional networks, encoder-decoder structure networks, multi-scale dilated convolutional neural networks, and attention mechanism-based networks are mainly used as the feature extraction module. These networks can generate prediction results for each pixel of the input image and achieve pixel-level semantic segmentation. However, the training data used in fully supervised semantic segmentation is costly and time-consuming and labor-intensive. As a result, algorithms based on weak supervision have emerged and been widely used in various fields. For example, Melih [2] applied it in the field of healthcare and obtained effective results. Most weakly supervised semantic segmentation approaches rooted in image-level labels typically consist of three stages. Firstly, a classification network is used to produce a preliminary localization map, referred to as a class activation map. Then the initial localization map is refined to produce pseudo-labels. Finally, these pseudo-labels are utilized to train a segmentation network. If the class activation map generated in the first stage is excessively sparse, it will significantly impact the performance of the subsequent two stages. Consequently, the research focus of this paper lies in determining how to ascertain methods for generating precise class activation maps.

2.2 CNN-Based Weakly Supervised Semantic Segmentation

Convolutional neural networks, as pivotal technologies in artificial intelligence, have found extensive applications across many fields with their advantages of diverse structural types and data-driven learning processes, especially in the field of image processing. For example, Zhang et al. [1] and others applied it to the field of image dehazing and achieved extremely excellent results. The field of image processing has witnessed rapid advancements in weakly supervised semantic segmentation, largely attributed to the evolution of convolutional neural networks. In the field of weakly supervised semantic segmentation, a common approach is to generate pseudo-labels using class activation maps generated by classification networks. These pseudo-labels are subsequently employed as training data for supervised segmentation tasks. However, a limitation of utilizing convolutional neural networks can lead to the problem of activating only local regions of the target. For this, many researchers have carried out research on the local problem of local activation problem in class activation maps and have continued to make significant progress in recent years. Kolesnikov et al. [18] extended the thermal range of CAM by enlarging the seed region. Wei et al. [19] activated the remaining lower discriminative regions by removing the higher discriminative regions of an object. Wang et al. [7] applied consistency regularization for the first time to CAM predicted from affine transformed images and proposed a pixel correlation module to enhance the consistency of CAM. Kumar et al. [20] randomly hid patches in the image during training as a way to motivate the network to explore other relevant parts. Zhang et al. [21] partitioned the image into complementary patches by in order to obtain a more comprehensive seed region in CAM. Ahn et al. [11] used a learning network to forecast the semantic affinity matrix among adjacent image coordinate pairs and used random walks to propagate the semantic affinity matrix. Despite numerous notable advancements in these research endeavors, methods rooted in convolutional neural networks (CNNs) still exhibit constraints in capturing global features, ultimately yielding suboptimal outcomes.

2.3 Transformer-Based Weakly Supervised Semantic Segmentation

Nowadays Vision Transformer has made significant breakthroughs in image processing tasks and many recent researches have applied it to the domain of weakly supervised semantic segmentation with good results. Researchers have proposed many advanced methods applied to WSSS by taking advantage of the Transformer's strengths in capturing global features. TS-CAM proposed by Gao et al. [22] provides an innovative method based on visual Transformer for weakly supervised object localization. This method utilizes the self-attention mechanism to achieve more extensive localization accuracy of the target area and significantly improves the localization accuracy. MCTformer+ proposed by Xu et al. [23] generates category-specific localization maps by introducing multiple class tokens and designs a contrastive class token module to enhance category discriminability. Zhu et al. [24] generate high-quality class activation maps through an adaptive attentional fusion module and introduce a gradient-clipping decoder for online retaining to improve segmentation accuracy.

2.4 Weakly Supervised Semantic Segmentation Based on the Fusion of CNN and Transformer

Since CNN and Transformer have their respective advantages when applied to WSSS tasks, some researchers have begun to try to utilize Conformer [17], which has a dual-branch structure, as a backbone network for generating more accurate class activation maps. The advantage of Conformer is to utilize CNN branches and Transformer components to interdependently fuse regional and comprehensive features as a way to simultaneously utilize the advantages of both branches in feature extraction. Based on the Conformer backbone network, Liu et al. [25] inspired by CPN [21] convert the input image into complementary patches. They reduce false detections by learning multiple estimation of complementary patches in different phases. At the same time, they also introduce an Adaptive Conflict Module (ACM) that can adaptively filter conflicting pixels and further improve the quality of pseudo-labels. He et al. [26] proposed a completely new adaptive reactivation mechanism, aiming to alleviate the uncontrolled over-smoothing problem of Transformer in weakly supervised semantic segmentation. This mechanism supervises the deep attention matrix to make it more focused on semantic objects, and further improves the quality of pseudo-labels while reducing background noise. Li et al. [16] proposed the TransCAM, which applies the Conformer as a backbone network. This network combines the attention weights generated by the Transformer component of the Conformer to enhance the class activation map produced by the CNN component, thus solving the constraint related to the limited local receptive field. However, due to the fact that Conformer's two-branch structure uses traditional convolutional blocks and Transformer blocks, it is not precise enough to capture different targets in the image and the computational complexity is too high. Therefore, this paper improves the main blocks of Conformer's dual-branch structure, and obtains an advanced method based on CNN and Transformer dual-branch structure.

3 Methodology

3.1 Network Architecture

To address the limitations of existing weakly supervised semantic segmentation methods, such as the loss of small target information, inability to handle complex and multi-scale target scenes, and high computational complexity of the network. We introduce a supervised semantic segmentation approach that leverages an enhanced Conformer network as its foundational backbone framework. The overall network architecture is shown in Fig. 1. First, the stem module is used to perform preliminary feature extraction on the image and input it to the CNN and Transformer branches. Then, the feature map of the last convolutional layer of the convolutional neural network is multiplied by the corresponding category weight and accumulated to obtain the class activation map. In this process, the coupling module is used to

continuously couple the features of the Transformer branch and the convolutional neural network branch. Finally, the attention weight from the Transformer component is used for further optimizing the class activation map produced by the CNN branch. To bolster the CNN branch's capability, we construct a cross-scale feature integration convolution module to capture more details and improve the sensitivity to multi-scale targets. For the Transformer branch, we construct an efficient multi-head self-attention module to increase the multi-scale feature extraction ability and reduce the computational complexity. Additionally, a multi-feature coupling module is designed to promote the improvement of the local-global feature mutual coupling performance when coupling the features of the dual-branch network.

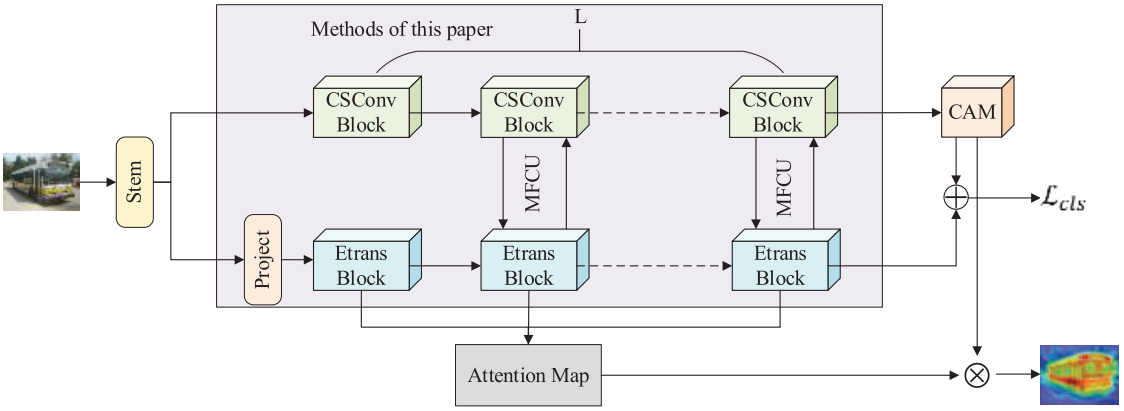

Figure 1: Multi-scale efficient weakly supervised semantic segmentation network

The overall network uses the same number of convolutional blocks and Transformer blocks to form a dual-branch structure to generate class activation maps, and uses the features of the Transformer branch to optimize the class activation maps. Our innovation lies in constructing a cross-scale feature integration convolutional module (CSConv Block) and an efficient multi-head self-attention module (Etrans Block) which are applied to the CNN and Transformer branches respectively, and constructing a multi-feature coupling module (MFUC) which is applied to the feature coupling of the two branches.

3.2 Cross-Scale Feature Integration Convolution Module

To enable the network to capture more details and improve the sensitivity to multi-scale targets, we construct a cross-scale feature integration convolutional module as the convolutional block of the CNN branch. This module can fuse multi-scale feature information and enhance the capacity to capture long-range pixel dependencies, thereby generating more comprehensive and accurate class activation maps. The specific structure is shown in Fig. 2.

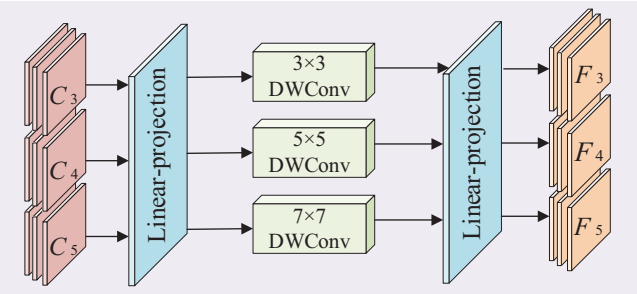

Figure 2: Cross-scale feature integration convolution module

The input image is initially extracted by the stem module in Fig. 1 and then obtains the feature pyramids C_3, C_4, and C_5 through a series of convolution operations. First, the dimension-reduced features are obtained through the linear projection layer. Then, the features are divided into M groups in the channel dimension and processed by depth wise separable convolutions with corresponding receptive field sizes. Finally, the processed features are connected and the dimension is increased through the linear projection layer.

After passing through the stem module in Fig. 1, the preliminary features of the image are obtained. Then, the feature map is further processed through a series of convolution operations. This generates a set of multi-scale features $\{C_3, C_4, C_5\}$ with resolutions of $\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$. These feature maps are then flattened and concatenated into a feature block $C \in R^{\left(\frac{HW}{8^2}+\frac{HW}{16^2}+\frac{HW}{32^2}\right)\times D}$, followed by dimensionality reduction through a linear projection layer. Next, the features are divided into M groups along the channel dimensions, and each group is processed by depthwise separable convolutions of different sizes to enhance the receptive field representation. Each group of features is processed by convolution layers with different receptive fields to increase receptive field diversity. Finally, the processed feature blocks are concatenated and expanded through another linear projection layer. The process of this module processing the feature map is shown in Eq. (1).

$$F = FC\left(DWConv\left(FC\left(C\right)\right)\right), \tag{1}$$

where F denotes the feature output by this module, $FC(\bullet)$ represents the linear projection,$DWConv(\bullet)$ denotes a set of depth wise convolutions with different convolution kernel sizes and C denotes the input multi-scale feature set. To retain multi-scale information while generating accurate class activation maps, we unify the feature map output from the last CNN block to the 7×7 resolution through feature fusion. This approach not only preserves multi-scale perception but also ensures that the final output feature map meets the required resolution through adjustment and fusion. Finally, the specific weight of the classification layer is multiplied by the corresponding channel of the feature map to obtain the activation map corresponding to each class.

3.3 Efficient Multi-Head Self-Attention Module

To compensate for the limitation of the CNN branch's local receptive field, a dual-branch network is formed by using the same number of Transformer blocks as the convolutional blocks. Starting from the second block, a coupling module is used for feature coupling between each group of convolutional blocks and Transformer blocks to achieve the complementary effect of local details and global features. Furthermore, the attention maps from each Transformer block are amalgamated into a single average attention map, which is then multiplied with the class activation map produced by the CNN part, yielding a more comprehensive

and precise class activation map. Since the self-attention mechanism in traditional Transformer blocks involves large matrix operations. When processing high-resolution images, this significantly increases the computational demand and complexity. To improve performance on images with different resolutions, we designed an efficient multi-head self-attention module. It replaces the traditional multi-head self-attention module in the Transformer block, aiming to reduce the computational load while maintaining image detail. The specific structure of the efficient multi-head self-attention module is shown in Fig. 3.

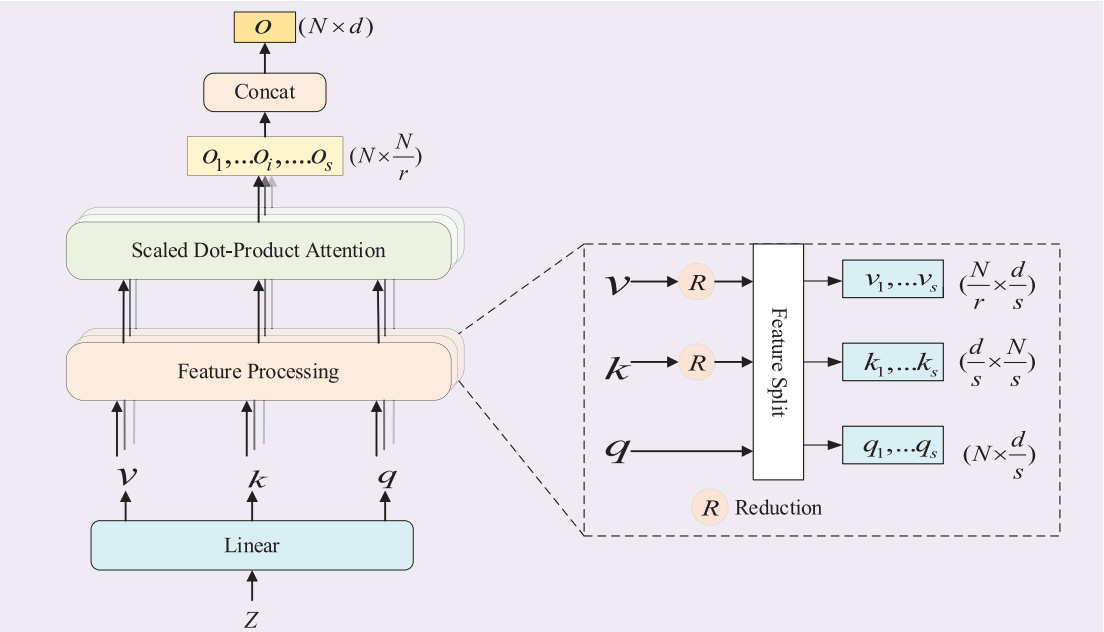

Figure 3: Efficient multi-head self-attention module

The input of this module is the D-dimensional latent embedding space Z generated by the project module in Fig. 1. First, q, k, and v of the ith head are obtained through linear projection. Then, the spatial resolution of k and v is reduced by a ratio of $r = 2$. Next, the sub-token matrix generated by the feature splitting operation is multiplied by the matrix representing the original perceptual region $\frac{1}{s}$. In this way, memory consumption can be effectively reduced.

This module reduces computational cost by reducing the feature dimension while trying to avoid the loss of image details as much as possible. We denote the input features of the project module in Fig. 1 as $X_t = R^{C_t \times H_t \times W_t}$, where C_t, H_t, W_t represents the channel, height and width of the feature map. First, after performing the Reshape operation in the project module, a flattened and non-overlapping patch sequence is obtained, resulting in $X_t \in R^{N \times (C_t \cdot P^2)}$, where $N = \frac{H_t W_t}{P^2}$ represents the number of patches (i.e., the length of the input sequence), and p^2 denotes the size of each patch. Then, these patches are mapped into a potential D-dimensional potential embedding space, through a learnable linear projection layer $E \in R^{(P^2 \cdot C_t) \times D}$, denoted as $Z \in R^{N \times D}$. This process is illustrated in Eq. (2).

$$Z = \left[x_p^1 E; x_p^2 E; \ldots x_p^N E\right], \tag{2}$$

where x_p^i represents the ith image block, and E represents a learnable linear projection layer. Next, it is input to the linear projection matrices W^Q, W^K and $W^V \in R^{D \times D_h}$, generating the queries Q, the keys K and the

values V in the self-attention mechanism. The detailed process is shown in Eq. (3).

$$Q, K, V = ZW^{Q}, ZW^{K}, ZW^{V} \in R^{N \times D_h}, \tag{3}$$

where N represents the count of patches, D denotes the embedding dimension of each patch, h represents the count of heads in the multi-head self-attention, a parameter specified by the user, which guarantees that the dimension of each head is $d = \frac{D_h}{h}$. Thus, after the above steps, in the ith head, the dimensions of q, k and v are $N \times d$. First, in the ith head, k and v will be spatially compressed with a compression factor of r $(r = 2)$. Then, the sub-token generated by feature splitting is matrix multiplied with a region that represents merely $\frac{1}{s}$ of the initial perception field, where s stands for the number of feature segmentations, set to 4 in this paper. the process is shown in Eq. (4).

$$(q_1, \ldots, q_s), (k_1, \ldots, k_s)(v_1, \ldots, v_s) = Feature\_Split(q, k, v), \tag{4}$$

where $q_i \in R^{N \times \frac{d}{s}}, k_i \in R^{\frac{N}{r} \times \frac{d}{s}}, v_i \in R^{\frac{N}{r} \times \frac{d}{s}}$. represents the segmented spatial distribution, and $Feature\_Split(q, k, v)$ represents the feature splitting operation. Through the above operations, the matrix is split into multiple sub-matrices along a specific dimension for subsequent grouped self-attention calculations, effectively reducing computational complexity and memory consumption after processing. The calculation of self-attention in the nth head is shown in Eqs. (5) and (6).

$$o_i(q_i, k_i, v_i) = Soft\ max\left(\frac{q_i(k_i)^T}{\sqrt{d}}\right)v_i, i \in [1, s], \tag{5}$$

$$head^n = Concat[o_1, o_{2,} \ldots o_s], n \in [1, h], \tag{6}$$

where o_i represents the self-attention output of the ith sub-token matrix, $Softmax(\cdot)$ represents applying the softmax function to the similarity to generate attention weights, $head^n$ represents the output of the n-th attention head, and $Concat[.,.]$ represents the concatenation operation. After the above operations, the ultimate output from the efficient multi-head self-attention module is derived, as shown in Eq. (7).

$$eMHSA = Concat[head^1, head^2, \ldots head^h]W^O, \tag{7}$$

where h denotes the count of heads in the efficient multi-head self-attention and $W^O \in R^{D_h \times D}$ is used as a linear projection to recover the dimension. By utilizing the efficient multi-head self-attention structure, the complexity of the network is reduced from $O\left(N^2\right)$ to $O\left(\frac{N^2}{sr}\right)$.

3.4 Multi-Feature Coupling Module

As can be known from Section 3.1, our network structure is a dual-branch network that concurrently utilizes the strengths of both CNN and Transformer. During the feature computation phase, it is necessary to continuously couple the features of the two branches to achieve the purpose of taking into account local and global features at the same time. Therefore, we have constructed a multi-feature coupling module to serve as the bridge connection part of the dual-branch network. The multi-feature coupling module is shown in Fig. 4.

The input of this module is the feature $F = \{F_2, F_3, F_4\} \in R^{\left(\frac{HW}{9^2} + \frac{HW}{16^2} + \frac{HW}{32^2}\right) \times D}$, outputted by the CNN branch and the feature $X \in R^{\frac{H}{16} \times \frac{W}{16} \times D}$, outputted by the Transformer branch. The feature obtained after the addition and self-attention unified operation can be used as an additional input for each module within the dual-branch network to make up for the original defects of the module.

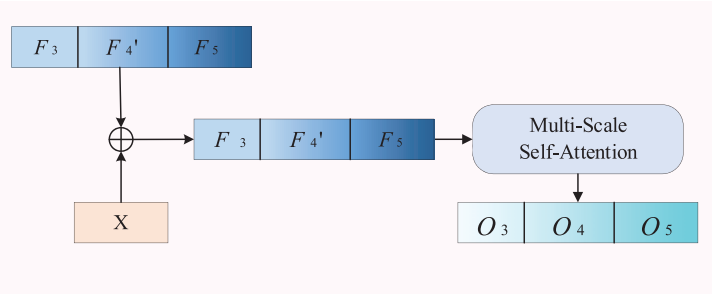

Figure 4: Multi-feature coupling module

First, the Transformer branch features X and the CNN branch feature with the same resolution F_3 are added together to obtain the fused feature F_3'. This fused feature set $F' = \{F_2, F_3', F_4\}$ combines multi-scale features from dual-branch network. To address the disparity in modal representations, a self-attention mechanism is applied to unify the features generated by the modules in the dual-branch network, reducing the impact of modal differences. This process is shown in Eq. (8).

$$O = FFN\left(Attention\left(norm\left(F'\right)\right)\right), \tag{8}$$

where F' includes multi-scale features with resolutions of $\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$, $norm\ (\bullet)$ denotes layer normalization [27]. $Attention\ (\bullet)$ denotes multiscale deformable attention [28], $FFN\ (\bullet)$ denotes feedforward network, and O represents the final feature output of this coupling module Bilinear interpolation is used to align the feature map sizes of O_3 and O_5 to O_4, and is fused with the outputs of the CNN module and Transformer block at this stage as the input for the modules in the next stage.

For the ViT branch, the features of the two branches are first coupled before the phase i begins, and the resulting of adding the coupled features to the output of the previous stage are then injected into the ViT branch. This process is illustrated in Eq. (9).

$$\hat{X}_i = \alpha * O_{i-1} + X_{i-1}, \tag{9}$$

where $\hat{X}_i$ denotes the updated features of the ViT branch the Transformer block of the ViT branch at stage i, α represents a learnable variable, O_{i-1} represents the coupling result of the dual-branch features of the previous stage, and X_{i-1} represents the output feature of the Transformer block of the previous stage Similarly, the feature update process for the CNN branch, the process is shown in Eq. (10).

$$\hat{F}_i = \alpha * O_{i-1} + F_{i-1}, \tag{10}$$

where $\hat{F}_i$ denotes the updated features from the CNN component, and F_{i-1} represents the output feature from the convolutional block of the previous stage.

After the above operations, a class activation map corresponding to the category in the image is obtained. Following this, the class activation map undergoes optimization to produce high-fidelity pseudo-labels, which are subsequently employed as the training dataset for semantic segmentation tasks.

4 Experimental Results and Discussions

4.1 Implementation Details

The experiment was conducted on a system with an Intel(R) Xeon(R) Silver 4110 CPU @ 2.10 GHz, 64 GB of memory, and an NVIDIA Quadro RTX 6000 GPU. The software environment consisted of Windows

10, Python 3.7, PyTorch 1.13, and Pycharm 2022.2.1. First, the proposed network was pre-trained on ImageNet. Then, fine-tuning was performed on the PASCAL VOC 2012 and MS COCO 2014 training sets using an NVIDIA RTX A6000 GPU (with 48 GB of memory). The AdamW [29] was used, with a learning rate set to $5e^{-5}$ and a weighting Decay set to $5e^{-4}$. Throughout the training process, the images underwent random resizing within the bounds of $[320, 640]$ and subsequently cropped to dimensions of 512 × 512. During inference, the images were input to the model are at size of 256 × 256, 512 × 512 and 768 × 768, generating CAMs at each scale. The CAMs are subsequently resized to match the dimensions of the initial image and integrated together. To enhance the prediction accuracy of mIoU, a multi-scale methodology is adopted.

4.2 Dataset and Evaluation Metric

In order to assess the datasets, we used the PASCAL VOC 2012 and MS COCO 2014 datasets. The PASCAL VOC 2012 dataset includes 21 categories, comprising 20 foreground categories and a single background class, and offers a dataset with 1464 training images, 1449 images for validation, and 1456 images for testing. Since the test set labels are not publicly available, model predictions must be submitted to the official website for performance evaluation. The MS COCO 2014 dataset consists of 1 background class and 80 foreground classes, with a total of 82,081 images designated for training and images for validation purposes. Following the standard experimental protocol used in previous studies [11,12,30], we extracted supplementary annotations from the semantic boundaries dataset [31], resulting in an expanded training set comprising 10,582 images.

During training, image-level class labels were used, and the performance of the proposed network architecture was assessed by the mean Intersection over Union (mIoU) metric. This measurement signifies the proportion of the overlapping area between the predicted and actual classes, relative to their combined area. A higher ratio indicates better performance. The calculation process is shown in Eq. (11).

$$mIoU = \frac{1}{C}\sum_{C}\frac{area\,(P) \cap area\,(G)}{area\,(P) \cup area\,(G)}, \tag{11}$$

where P denotes the prediction result, G denotes the real label, $area\,(\cdot)$ represents the area of the corresponding region, and C represents the count of target categories in the dataset.

4.3 Ablation Studies

In order to estimate the influence of each distinct module upon the network's operational efficiency, the TransCAM architecture was used as the baseline. The dual-branch structure and fusion module were progressively replaced to assess the effectiveness of the cross-scale feature integration convolution module, efficient multi-head self-attention module, and multi-feature coupling module. Through these replacements, the impact of each module on the mIoU was evaluated on the PASCAL VOC 2012 training set. The results are shown in Table 1.

The "$\surd$" symbol in the table indicates that the module was used, while the absence of this symbol indicates that the module was not used. As shown in Table 1, the baseline model achieved a mIoU of 64.3%. This relatively low value is primarily attributed to the single-scale feature maps. After replacing the traditional multi-head self-attention in ViT with the proposed efficient multi-head self-attention module, the mIoU of the generated class activation maps is 64.2%, indicating no significant impact on the network's performance. However, when the proposed cross-scale feature integration CNN module was introduced into the CNN branch, combined with the multi-feature coupling module, the mIoU increased to 72.8%, representing a 13.2% improvement over the baseline network. The experiment clearly demonstrates that the proposed modules can effectively enhance the network's performance.

Table 1: Results of ablation experiments with different modules

Standard	Cross-scale feature integration convolution module	Efficient multi-head self-attention module	Multi-feature coupling modula	mIoU/%↓
√				64.3
√		√		64.2
√	√	√	√	**72.8**

To evaluate the impact of the efficient multi-head self-attention module on network performance, we used the ViT network as the baseline and compared the effects on segmentation accuracy and computational complexity before and after replacing the traditional multi-head self-attention module. The results are shown in Table 2.

Table 2: Results of our proposed structure for performance improvement

Standard	Efficient multi-head self-attention module	Param (K)↓	FLOPs (G)↓	mIoU/%↓
√		680.352	13.294	64.3
√	√	569.937	11.581	64.2

The "√" symbol in the table indicates that the module was used, while the absence of this symbol indicates that the module was not used. The Param in Table 2 represents the number of model parameters. Fewer parameters indicate lower model complexity, meaning the model requires less memory and computational resources during training and inference. FLOPs represents the number of floating-point operations performed per forward pass, measured in gigaflops (G). A lower FLOPs value suggests a more efficient model, which may result in faster inference speeds. The results show that the efficient multi-head self-attention module strikes a favorable equilibrium between performance efficiency and precision. Compared to the baseline structure, the proposed framework reduces the number of parameters by 24% and decreases the computational load by 13%, with only a slight 0.2% drop in mIoU. This indicates that the efficient multi-head self-attention module significantly enhances network efficiency and accelerates inference speed with minimal impact on performance.

4.4 Comparisons with SOTA Methods

To enhance the mIoU of pixel-level pseudo-labels, PSA [11] was first used for post-processing, followed by dense CRF [32] to further refine the previously generated pseudo-labels. To evaluate the potency of our method in generating class activation maps, we compared the generated CAMs with recent research results, as shown in Fig. 5.

The generated CAMs were used as pseudo-labels to train the classic segmentation model DeepLab [33] under full supervision, with ResNet38 [34] as the backbone network. To assess the efficacy of the proposed method further, we compared its segmentation outcomes with those of existing algorithms on the PASCAL VOC 2012 and MS COCO datasets, as presented in Tables 3 and 4.

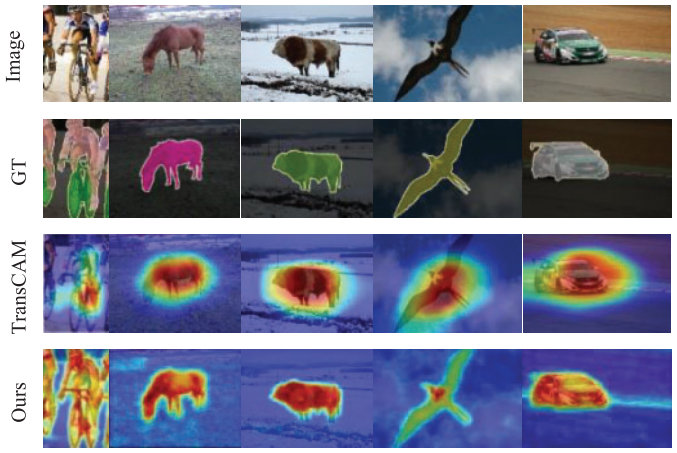

Figure 5: Results of the comparison of CAM visualizations

Table 3: Results of segmentation performance comparison on PASCAL VOC dataset

Method	Publication	Sup.	Val	Test
AuxSegNet [35]	ICCV21	I + S	69.0	68.6
L2G [36]	CVPR22	I + S	72.0	73.0
MECPformer [25]	Arxiv23	I + S	72.0	72.0
SEAM [7]	CVPR20	I	64.5	65.7
CONTA [37]	NIPS20	I	66.1	66.7
CDA [38]	ICCV21	I	66.1	66.8
CPN [22]	ICCV21	I	67.8	68.5
AdvCAM [13]	CVPR22	I	68.1	68.0
AMN [39]	CVPR22	I	70.7	70.6
W-OoD [3]	CVPR22	I	70.7	70.1
SIPE [40]	ECCV22	I	68.2	69.5
Yoon et al. [41]	CVPR23	I	70.9	71.7
OCR [42]	CVPR23	I	72.7	70.7
LPCAM [43]	CVPR23	I	72.6	72.4
MCTformer [23]	Arxiv23	I	74.0	73.6
He et al. [26]	Arxiv23	I	69.9	70.0
TransCAM [16]	JVCI23	I	69.3	69.6
Ours	–	I	71.3	72.1

Table 4: Results of segmentation performance comparison on MS COCO dataset

Method	Publication	Sup.	Val
AuxSegNet [35]	ICCV21	I + S	33.9
L2G [36]	CVPR22	I + S	44.2
MECPformer [25]	Arxiv23	I + S	42.4
SEAM [7]	CVPR20	I	31.9

(Continued)

Table 4 (continued)

Method	Publication	Sup.	Val
CONTA [37]	NIPS20	I	32.8
CDA [38]	ICCV21	I	33.2
AdvCAM [13]	CVPR22	I	44.4
AMN [39]	CVPR22	I	44.7
SIPE [40]	ECCV22	I	43.6
Yoon et al. [41]	CVPR23	I	44.8
OCR [42]	CVPR23	I	42.5
LPCAM [43]	CVPR23	I	42.8
MCTformer [23]	Arxiv23	I	45.2
TransCAM [16]	JVCI23	I	45.7
Ours	–	I	46.3

In Tables 3 and 4, the symbol I represents image-level labels, whereas S signifies significance maps. As presented in Tables 3 and 4, the proposed approach attains mIoU scores of 71.3% and 72.1% on the PASCAL VOC 2012 validation and test datasets, and achieves 46.3% on the MS COCO validation dataset. The results demonstrate that the segmentation outcomes achieved by the proposed method surpass those of alternative approaches that solely rely on image-level annotations, by a notable margin. To a more intuitive demonstration of the suggested approach's performance, the semantic segmentation results are visualized in Fig. 6, where (a) depicts the initial image, (b) displays the ground truth, and (c) shows the predictions produced by our method.

To further validate the potency of our proposed approach on specific categories, we conducted a comparative analysis using several recent algorithms on the validation dataset from PASCAL VOC 2012. The performance of each method was evaluated with the Intersection over Union (IoU) for each category and the mIoU across all categories. The outcomes of the comparison are presented in Table 5.

The mIoU for all categories is shown in the last column of Table 5, with the best-performing algorithm for each category highlighted in bold. Our proposed method exhibits superior performance in 16 out of the 21 categories, indicating the efficacy of the designed network across both specific and individual categories, as compared to existing approaches. In terms of overall segmentation performance, the proposed method surpasses the current state-of-the-art algorithms.

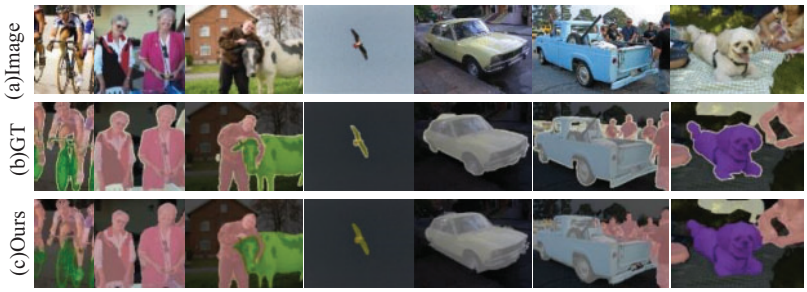

Figure 6: Visualization of segmentation results

Table 5: Results of segmentation performance comparison on PASCAL VOC dataset

Method	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow
SEC [18]	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	75.2	22.1	53.5
PSA [11]	89.1	70.6	31.6	77.2	42.2	**68.9**	79.1	66.5	74.9	29.6	68.7
FickleNet [12]	90.3	77.0	35.2	76.0	54.2	64.3	76.6	76.1	80.2	25.7	68.6
RRM [44]	87.9	75.9	31.7	78.3	54.6	62.2	80.5	73.7	71.2	30.5	67.4
TransCAM [16]	91.3	81.9	35.4	84.7	67.6	67.9	87.5	80.5	86.5	**31.4**	73.9
Ours	**91.6**	**85.3**	**37.5**	**85.1**	**69.0**	68.2	**89.2**	**81.7**	**88.9**	29.4	**79.2**
Method	**table**	**dog**	**horse**	**mbk**	**person**	**plant**	**sheep**	**sofa**	**train**	**tv**	**mIoU**
SEC [18]	28.3	65.8	57.8	62.3	52.5	32.5	62.6	32.1	45.4	45.3	50.7
PSA [11]	56.1	82.1	64.8	78.6	73.5	50.8	70.7	47.7	63.9	51.1	63.7
FickleNet [12]	50.2	74.6	71.8	78.3	69.5	**53.8**	76.5	41.8	70.0	54.2	65.3
RRM [44]	40.9	71.8	66.2	70.3	72.6	49.0	70.7	38.4	62.7	**58.4**	62.6
TransCAM [16]	52.5	**84.0**	74.9	74.6	79.0	44.7	84.1	47.0	78.4	46.6	69.3
ours	**56.7**	83.1	**79.8**	**81.9**	**79.4**	53.2	**85.4**	**48.1**	**79.1**	55.6	**71.8**

5 Conclusions

Through this research, we propose a multi-scale semantic segmentation network based on an improved Conformer. The proposed method can accurately capture objects of different sizes, enhance sensitivity to image details, and reduce computational complexity of multi-head self-attention in the Transformer. We use a dual-branch network of CNN and Transformer for local-global feature complementation to produce a more comprehensive class activation map, and use attention features from the Transformer to further refine the resulting class activation map. A cross-scale feature integration convolution module was designed for the CNN branch, incorporating multi-receptive field convolution layers to bolster the model's capacity to grasp long-range dependencies, thus enhancing its efficacy in managing objects of diverse sizes. For the Transformer branch, an efficient multi-head self-attention module was developed, applying spatial compression to the keys and values to minimize the transformer blocks' overall computational intricacy. Finally, a multi-feature coupling module was constructed to fully leverage the strengths of both CNN and Transformer. Extensive experimental results demonstrate that the average intersection over union (IoU) index of the image semantic segmentation results of the method in this paper on the validation set and test set of the PASCAL VOC 2012 dataset is 2.9% and 3.6% higher than that of the TransCAM algorithm respectively; on the COCO 2014 dataset, the average intersection over union index (mIoU) is 1.3% higher than that of the TransCAM algorithm. It is superior to the existing mainstream algorithms. The parameter amount index

Param. and the number of operations index FLOPs of the improved model have decreased by 16.2% and 12.9%, respectively. Future improvements can be made based on the dual-branch architecture to address the issue of still having suboptimal effects when dealing with complex scenes.

Acknowledgement: The authors would like to express their gratitude to all the researchers and reviewers who contributed to enhancing the quality of the idea, concept, and the paper overall.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Xueli Shen, Meng Wang; data collection: Meng Wang; analysis and interpretation of results: Xueli Shen, Meng Wang; draft manuscript preparation: Xueli Shen, Meng Wang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data presented in this study are available on request from the corresponding author. Data are not publicly available due to privacy considerations.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Zhang H, Wei Y, Zhou H, Wu Q. ED-Dehaze Net: encoder and decoder dehaze network. Int J Interact Multi. 2022 Sep;7(5):93. doi:10.9781/ijimai.2022.08.008.
2. Li X, Zhang J, Yang Y, Cheng G, Yang K, Tong Y, et al. Sfnet: faster and accurate semantic segmentation via semantic flow. Int J Comput Vis. 2024 Sep;132(2):466–89. doi:10.1007/s11263-023-01875-x.
3. Wei Z, Chen L, Jin Y, Ma X, Liu T, Ling P, et al. Stronger fewer & superior: harnessing vision foundation models for domain generalized semantic segmentation. Paper presented at: 2024 IEEE/CVF conference computer vision and pattern recognition; 2024 Jun 17–21; Seattle, WA, USA. doi:10.1109/CVPR52733.2024.02704.
4. Lee J, Yi J, Shin C, Yoon S. BBAM: bounding box attribution map for weakly supervised semantic and instance segmentation. Paper presented at: the IEEE/CVF conference computer vision and pattern recognition; 2021 Jun 10–25; Nashville, TN, USA. doi:10.1109/cvpr46437.2021.00267.
5. Lin D, Dai J, Jia J, He K, Sun J. ScribbleSup: scribble-supervised convolutional networks for semantic segmentation. Paper presented at: the IEEE conference computer vision and pattern recognition; 2016 Jun 27–30; Las Vegas, NV, USA. doi:10.1109/CVPR.2016.344.
6. Bearman A, Russakovsky O, Ferrari V, Fei-Fei L. What's the point: semantic segmentation with point supervision. Paper presented at: the european conference on computer vision; 2016 Oct 11–14; Amsterdam, The Netherlands. doi:10.1007/978-3-319-46478-7_34.
7. Wang Y, Zhang J, Kan M, Shan S, Chen X. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. Paper presented at: The IEEE/CVF conference computer vision pattern recognition; 2020 Jun 13–19; Seattle, WA, USA. doi:10.1109/CVPR42600.2020.01229.
8. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis. 2015 Apr;115(3):211–52. doi:10.1007/s11263-015-0816-y.
9. Melih A. Semi-supervised machine learning approaches for thyroid disease prediction and its integration with the internet of everything. Int J Interact Multi. 2024 Jul;8(7):38. doi:10.9781/ijimai.2024.07.006.
10. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. Paper presented at: the IEEE conference on computer vision pattern recognition; 2016 Jun 27–30; Las Vegas, NV, USA. doi:10.1109/CVPR.2016.319.

11. Ahn J, Wak S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. Paper presented at: the IEEE conference on computer vision pattern recognition; 2018 Jun 18–22; Salt Lake City, UT, USA. doi:10.1109/cvpr.2018.00523.
12. Lee J, Kim E, Lee S, Lee J, Yoon S. FickleNet: weakly and semi-supervised semantic image seg-mentation using stochastic inference. Paper presented at: the IEEE/CVF conference on computer vision pattern recognition; 2019 Jun 15–20; Long Beach, CA, USA. doi:10.1109/CVPR.2019.00541.
13. Lee J, Kim E, Yoon S. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. Paper presented at: the IEEE/CVF conference on computer vision pattern recognition; 2021 Jun 10–25; Nashville, TN, USA. doi:10.1109/cvpr46437.2021.00406.
14. Alexey D. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
15. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Paper presented at: the advances in neural information processing systems 30 (NIPS 2017); 2017 Dec 4–9; Los Angeles, CA, USA. doi:10.7312/burn21118-012.
16. Li R, Mai Z, Zhang Z, Jang J, Sanner S. TransCAM: transformer attention-based CAM refinement for weakly supervised semantic segmentation. J Vis Commun Image Rep. 2023 Apr;92(4):103800. doi:10.1016/j.jvcir.2023.103800.
17. Peng Z, Huang W, Gu S, Xie L, Wang Y, Jiao J, et al. Conformer: local features coupling global representations for visual recognition. Paper presented at: the IEEE/CVF international conference on computer vision; 2021 Oct 10–17; Montreal, QC, Canada. doi:10.1109/iccv48922.2021.00042.
18. Kolesnikov A, Lampert CH. Seed, expand and constrain: three principles for weakly-supervised image segmentation. Paper presented at: the computer vision–ECCV 2016: 14th european conference; 2016 Oct 11–14; Amsterdam, The Netherlands. doi:10.1007/978-3-319-46493-0_42.
19. Wei Y, Feng J, Liang X, Cheng MM, Zhao Y, Yan S. Object region mining with adversarial erasing: a simple classification to semantic segmentation approach. Paper presented at: the IEEE conference on computer vision pattern recognition; 2017 Jun 21–26; Honolulu, HI, USA. doi:10.1109/CVPR.2017.687.
20. Kumar SK, Jae LY. Hide-and-seek: forcing a network to be meticulous for weakly-supervised object and action localization. Paper presented at: proceedings of the IEEE international conference on computer vision; 2017 Oct 22–29; Venice, Italy. doi:10.1109/iccv.2017.381.
21. Zhang F, Gu C, Zhang C, Dai Y. Complementary patch for weakly supervised semantic segmentation. In: Paper presented at: proceedings of the IEEE/CVF international conference on computer vision; 2021 Oct 10–17; Montreal, QC, Canada. doi:10.1109/iccv48922.2021.00715.
22. Gao W, Wan F, Pan X, Peng Z, Tian Q, Han Z, et al. TS-CAM: token semantic coupled attention map for weakly supervised object localization. Paper presented at: the proceedings of the IEEE/CVF international conference vision on computer; 2021 Oct 10–17; Montreal, QC, Canada. doi:10.1109/iccv48922.2021.00288.
23. Xu L, Bennamoun M, Boussaid F, Laga H, Ouyang W, Xu D. MCTformer: multi-class token transformer for weakly supervised semantic segmentation. IEEE Trans Pattern Anal Mach Intell. 2024 May;46:1–16. doi:10.1109/TPAMI.2024.3404422.
24. Zhu L, Li Y, Fang J, Liu Y, Xin H, Liu W, et~al. WeakTr: exploring plain vision transformer for weakly-supervised semantic segmentation. arXiv:2304.01184. 2023.
25. Liu C, Li G, Shen Y, Wang R. MECPformer: multi-estimations complementary patch with CNN-transformers for weakly supervised semantic segmentation. Neural Comput. Appl. 2023 Sep;35(31):23249–64. doi:10.1007/s00521-023-08816-2.
26. He J, Cheng L, Fang C, Zhang D, Wang Z, Chen W. Mitigating undisciplined over-smoothing in transformer for weakly supervised semantic segmentation. arXiv:2305.03112. 2023.
27. Ba JL. Layer normalization. 2016. arXiv:1607.06450.
28. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: deformable transformers for end- to-end object detection. arXiv:2010.04159. 2020.
29. Loshchilov I. Decoupled weight decay regularization. arXiv:1711.05101. 2017.

30. Jo S, Yu IJ. Puzzle-CAM: improved localization via matching partial and full features. Paper presented at: 2021 IEEE international conference image process (ICIP); 2021 Sep 19–22; Anchorage, AK, USA. doi:10.1109/icip42928.2021.9506058.
31. Hariharan B, Arbeláez P, Bourdev L, Maji S, Malik J. Semantic contours from inverse detectors. Paper presented at: the 2011 international conference on computer vision; 2011 Nov 6–13; Barcelona, Spain. doi:10.1109/iccv.2011.6126343.
32. Krähenbühl P, Koltun V. Efficient inference in fully connected crfs with gaussian edge potentials. Paper presented at: the advances in neural information processing systems; 2011 Dec 12–17; Granada, Spain. doi:10.1109/cvpr.2012.6247724.
33. Chen LC. Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv:1412.7062. 2014.
34. Wu Z, Shen C, Van DA. Wider or deeper: revisiting the resnet model for visual recognition. Pattern Recognit. 2019 Jun;90(3):119–33. doi:10.1016/j.patcog.2019.01.006.
35. Xu L, Ouyang W, Bennamoun M, Boussaid F, Sohel F, Xu D. Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. Paper presented at: the proceedings of the IEEE/CVF international conference on computer vision; 2021 Oct 10–17; Montreal, QC, Canada. doi:10.1109/iccv48922.2021.00690.
36. Jiang PT, Yang Y, Hou Q, Wei Y. L2G: a simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. Paper presented at: the proceedings of the IEEE/CVF conference on computer vision pattern recognition; 2022 Jun 18–24; New Orleans, LA, USA. doi:10.1109/cvpr52688.2022.01638.
37. Zhang D, Zhang H, Tang J, Hua XS, Sun Q. Causal intervention for weakly-supervised semantic segmentation. Adv Neural Inf Process Syst. 2020;33:655–66.
38. Su Y, Sun R, Lin G, Wu Q. Context decoupling augmentation for weakly supervised semantic segmentation. Paper presented at: the proceedings of the IEEE/CVF international conference on computer vision; 2021 Oct 10–17; Montreal, QC, Canada.
39. Lee M, Kim D, Shim H. Threshold matters in WSSS: manipulating the activation for the robust and accurate segmentation model against thresholds. Paper presented at: the proceedings of the IEEE/CVF conference on computer vision pattern recognition; 2022 Jun 18–24; New Orleans, LA, USA. doi:10.1109/cvpr52688.2022.00429.
40. Chen Q, Yang L, Lai JH, Xie X. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. Paper presented at: the proceedings of the IEEE/CVF conference on computer vision pattern recognition; 2022 Jun 18–24; New Orleans, LA, USA. doi:10.1109/cvpr52688.2022.00425.
41. Yoon SH, Kweon H, Cho J, Kim S, Yoon KJ. Adversarial erasing framework via triplet with gated pyramid pooling layer for weakly supervised semantic segmentation. Paper presented at: the european conference on computer vision; 2022 Oct 23–27; Tel Aviv, Israel. doi:10.1007/978-3-031-19818-2_19.
42. Cheng Z, Qiao P, Li K, Li S, Wei P, Ji X, et al. Out-of-candidate rectification for weakly supervised semantic segmentation. Paper presented at: the proceedings of the IEEE/CVF conference on computer vision pattern recognition; 2023 Jun 18–22; Vancouver, BC, Canada. doi:10.1109/cvpr52729.2023.02267.
43. Chen Z, Sun Q. Extracting class activation maps from non-discriminative features as well. presented at the IEEE/CVF Conference on Computer vision pattern recognition; 2023 Jun 18–22; Vancouver, BC, Canada. doi:10.1109/cvpr52729.2023.00306.
44. Zhang B, Xiao J, Wei Y, Huang K, Luo S, Zhao Y. End-to-end weakly supervised semantic segmentation with reliable region mining. Pattern Recognit. 2022 Aug;128:108663. doi:10.1016/j.patcog.2022.108663.