ISSN 2616-5775 Vol. 7, Issue 3: 21-31, DOI: 10.25236/AJCIS.2024.070303

Bi-Branch Weakly Supervised Semantic Segmentation with Transformer

Yijiang Wang[1,a,*], Hongxu Zhang[2,b]

[1]*School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China*
[2]*School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China*
[a]*wyj@home.hpu.edu.cn*, [b]*zhx@home.hpu.edu.cn*
[*]*Corresponding author*

Abstract: *Weakly supervised semantic segmentation (WSSS) based on image-level labels has garnered widespread attention due to its cost-effectiveness. In image-level WSSS, existing methods typically rely on CNNs to generate Class Activation Mapping (CAM) for locating object regions and obtaining pseudo labels. However, CAM often focus solely on discriminative regions, neglecting other valuable information in each image and resulting in incomplete localization maps. To address the partial activation issue of CAM, we propose a* ***Bi****-Branch Weakly Supervised Semantic Segmentation with* ***Trans****former (Bi-Trans) approach, which includes class-specific seed (SC-CAM) generation and consistency loss (SCC Loss), as well as pairwise affinity consistency loss (PAC Loss). Specifically, the initial seeds of class-specific are directly extracted using the Multi-Head Self-Attention (MHSA) mechanism in the Transformer encoder, bypassing the need for complex training. The SCC Loss aims to minimize the distance between initial seeds generated from two different views, thereby enhancing the feature representation of the original seeds and improving their quality. Additionally, the PAC Loss ensures consistency in regional affinity within each view, enhances target similarity in the affinity matrix, and effectively mitigates background noise issues in the seed region. We evaluate our method on the PASCAL VOC and 2012 MS COCO 2014 segmentation benchmarks. The results demonstrate that our Bi-Trans approach produces superior pseudo-masks using only image-level labels, achieving improved WSSS performance.*

Keywords: *WSSS, CAM, Transformer, SCC Loss, PAC Loss*

1. Introduction

Semantic segmentation is an important task in computer vision, which aims to assign each pixel in an image to its corresponding semantic class. Semantic segmentation based on deep learning is currently one of the most popular methods. However, this semantic segmentation method usually relies on much well-labeled data to achieve excellent results. In order to deal with this problem, weakly supervised semantic segmentation has gradually become a direction that has attracted much attention from researchers. Weakly supervised semantic segmentation means that during the training process, only coarse supervision (including image-level[1], point[2], scribbles[3], and bounding box[4]) is proactively employed to mitigate the reliance on pixel-level real labels, which are comparatively expensive. Coarse supervised annotations are cheaper than these, and large-scale image-level labels are common in the network and easier to obtain. Therefore, research based on image-level weakly supervised semantic segmentation (WSSS) has high theoretical research value.

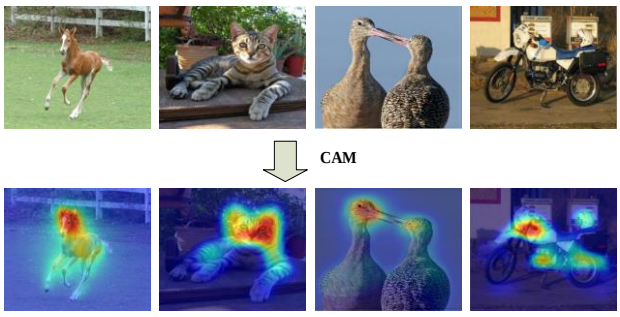

Figure 1: Example of CAM generated by CNN.

ISSN 2616-5775 Vol. 7, Issue 3: 21-31, DOI: 10.25236/AJCIS.2024.070303

However, the weakly supervised semantic segmentation task based on image-level labels is the most challenging because it uses class labels to supervise pixel-level predictions, only indicating the presence or absence of certain classes without actual location, shape, and size information about the object class. In other words, no direct mapping between semantic labels and pixels exists. Previous weakly supervised semantic segmentation methods usually rely on the Class Activation Mapping (CAM)[5] method based on convolutional neural networks to generate a coarse initial class localization map, which we call a seed, which only focuses on a small and precise target area (we usually refer to it as the discriminative area, as shown in Figure 1), such as the "head" of "cat," the "mouth" of "bird," etc. Despite using sophisticated CAM seed mining strategies or multi-network parallel training steps, the initial seeds generated by existing methods still need to be completed and accurate activations. The gap between image-level labeling and pixel-by-pixel segmentation supervision causes these problems.

The Vision Transformer[6] model has achieved performance breakthroughs in multiple visual tasks in recent years with its excellent global context modeling capabilities. Even in DINO[7], the Vision Transformer network model has been proven to have good segmentation potential. Thanks to Transformer's excellent global feature interaction capabilities, it can solve the problem that the convolutional neural network model only focuses on the discriminative area and generates low-quality initial seeds. This paper proposes **Bi**-Branch Weakly Supervised Semantic Segmentation with Transformer (Bi-Trans). It is worth noting that our backbone network uses the multi-class token DeiT Transformer model. Like DINO, our backbone network is also a novel knowledge distillation network. The difference is that we use multi-class toekn[8] for a specific weakly supervised semantic segmentation task to directly generate class-specific initial seeds using self-attention maps. On the one hand, inspired by the convolutional neural network method[9], a self-supervised Bi-branch Transformer network is proposed. Performing an affine transformation on the original image proactively yields the enhanced view.The two views are jointly input into the network to obtain two view class-specific attention matrices, proactively using it as the initial seed. The distance between the initial seeds generated by two different views is shortened through the proposed specific class consistency loss (SCC Loss), minimizing the difference between the two, thereby enhancing the feature representation of the seed and ultimately improving the quality of the initial seed. However, we found during the experiment that the Transformer's long-range context extraction capability inevitably brings significant background noise to the seeds. However, the pairwise attention matrix between all patch markers learned through network training can use as a pairwise affinity matrix, which proactively refines the generated initial seeds to mitigate the impact of background noise.Therefore, we propose the pairwise affinity consistency loss (PAC Loss) to narrow the gap between different affinity matrices obtained through different views and enhance the consistency between affinities. This approach allows us to address the background noise problem in the network's initial seeds while obtaining higher-quality pseudo masks.

Our main contributions are summarized as:

- We propose Bi-Branch Weakly Supervised Semantic Segmentation with Transformer (Bi-Trans) to provide self-supervised information for weakly supervised semantic segmentation network learning.
- We propose the Specific Class Consistency (SCC) loss to effectively regularize the initial seeds generated by two views and minimize the difference between them.
- We propose Pairwise Affinity Consistency (PAC) loss, which forces features to be consistent between different affinity matrices obtained from different views, enhances the target similarity in the affinity matrix, and solves the background noise problem in the initial seeds generated by the network.

2. Proposed Method

This section introduces the implementation details of this algorithm in detail. First, the detailed modules included in the entire dual-branch network and the processing pipeline of the algorithm are described. Secondly, we introduce the implementation of class-specific Seed Generation (SC-CAM) and Specific Class Consistency (SCC Loss). Principles. Finally, the PAC loss regularization method is described to enhance the consistency between the two affinity matrices, using the affinity matrix of the original image to refine the generated initial seed.

2.1. Overview

Inspired by traditional weakly supervised semantic segmentation work[9], generating an initial class activation map for the affine transformed image can provide self-supervision for the original network

ISSN 2616-5775 Vol. 7, Issue 3: 21-31, DOI: 10.25236/AJCIS.2024.070303

learning. Compared with Convolutional Neural Networks (CNNs), the Transformer explicitly encodes the regional dependencies between all markers through a self-attention mechanism. These characteristics align well with our need to model the feature consistency of two input forms without introducing additional convolution modules and complex loss constraints, which differs from existing methods based on traditional convolutional neural networks. Therefore, this paper designs a Bi-Branch Weakly Supervised Semantic Segmentation with Transformer, Bi-Trans. The network structure is shown in Figure 2.

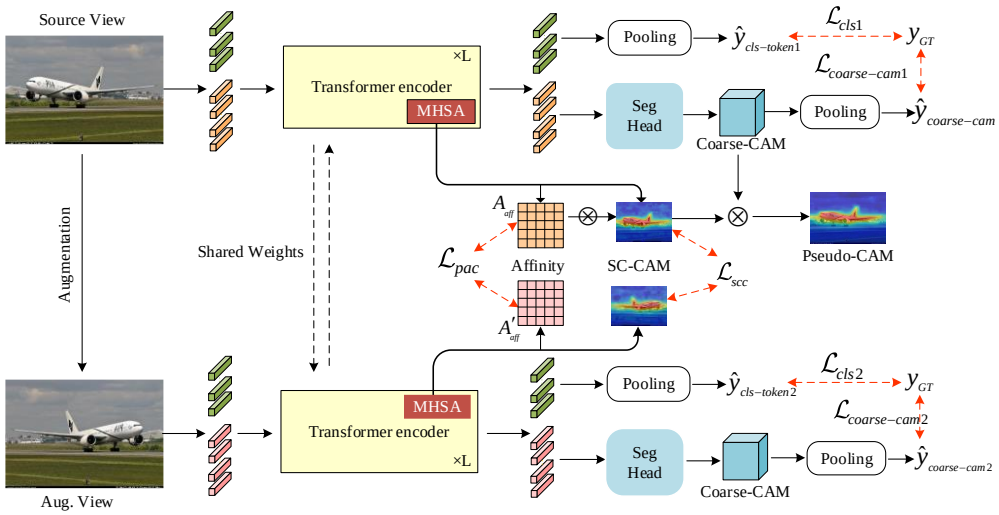

Figure 2: Overview of the proposed Bi-Trans for weakly supervised semantic segmentation.

Specifically, the dual-branch network proposed in this article is a twin network, which contains two branches comprising a pair of input images (an original image and another data-augmented image) to learn class-specific Attention maps generated self-supervised. Each branch contains a backbone network DeiT with Transformer encoders, in which the Transformer encoder inputs class tokens corresponding to the number of classes in different data sets and participates in the network training to obtain the attention map for class-specific. The network obtains different specific-class initial seeds and pair-wise affinity matrices through two input images. In addition, we also add a segmentation head (Seg head) after the output patch tokens of the Transformer to generate a coarse but highly responsive initial pseudo mask (Coarse-CAM), which can assist in the refinement of the initial seed together with the pair-wise affinity matrix. Finally, our proposed pair-wise affinity consistency loss constrains the pair-wise affinity matrix generated by different input images. It ultimately utilizes the refined affinity matrix to enhance the initial seed. The entire network requires no additional training modules. Our method demonstrates higher training efficiency and faster inference speed than convolutional neural network-based methods.

$$X' = HorizontalFlip(X) \tag{1}$$

Notably, the data enhancement method used in this paper is horizontal flipping, and another input image is obtained through horizontal flipping, as shown in Equation (1). This differs from the traditional CNN method that uses scaling as a data enhancement method. Considering the inherent characteristics of Vision Transformer, when the input image size changes and the patch tokens size remains unchanged, its number will change, positively correlated with the size change. As a result, the patch embedding size is inconsistent with the patch embedding size of the original image. Therefore, the scaling method is not conducive to our subsequent calculation of specific-class affinity consistency loss. For convenience, we chose a data enhancement method that does not change the number of patches, flip, and rotation. There are two flip methods in Pytorch: horizontal flip and vertical flip. Through subsequent experimental comparisons, it was found that the seed performance is best when the horizontal flip is used as a data enhancement method.

2.2. Class-specific Seed generation and Consistency Loss

The main work of this section is to use a Bi-branch network architecture to generate initial seeds of different views for weakly supervised semantic segmentation tasks, laying the foundation for subsequent pseudo masks for fully supervised semantic segmentation tasks. Then, through the class-specific consistency loss proposed in this paper, the gap between the initial seeds of different views is reduced,

ISSN 2616-5775 Vol. 7, Issue 3: 21-31, DOI: 10.25236/AJCIS.2024.070303

the consistency between the two is enhanced, and the seeds of the original image are forced to have higher activation in the foreground.

For the original input image, divide it into $N \times N$ patches, flatten them, and linearly map them into $N \times N$ patch tokens $T_p \in \mathbb{R}^{M \times D}$, where D is the embedding dimension, $M = N^2$. Further, C learnable class tokens are generated, where C represents the total number of classes in the data set, concatenated with the patch tokens as the input $T_{in} \in \mathbb{R}^{(C+M) \times D}$ of the Transformer encoder. The Transformer encoder has L consecutive encoder layers. Each encoder layer consists of a Multi-Head Attention (MHA) module, a Multi-layer Perceptron (MLP), and the LayerNorm (LN) layer. In each encoder layer, we input tokens T_{in} and receive T_{out}, which becomes the new T_{in} for the next encoder layer, and so on, for a total of L iterations.

This paper uses a standard self-attention mechanism to capture long-distance dependencies between tokens. We obtain the attention map a between all pairs of tokens, from which we can extract the class-specific initial seed SC-CAM we need, and b, the class-to-image patch attention map. The SC-CAM calculation is shown in Equation (2):

$$M = A_{att}[1:C, C:C+N^2] \tag{2}$$

Similarly, the augmented view branch also gets the initial seed SC-CAM, $\hat{M} \in \mathbb{R}^{C \times N^2}$.

To further exploit class-specific knowledge, this paper introduces a self-supervised learning paradigm: Class-Specific Consistency Loss (SCC Loss, $\mathcal{L}_{scc}$). The $\mathcal{L}_{scc}$ is a regularization loss used to minimize the difference between the initial seed SC-CAM generated from the original image and the SC-CAM of the seed generated from the data-augmented image. The mathematical definition of this consistency regularization is formulated as the L1 regularization of two class-specific CAMs, also known as the L1 norm. The $\mathcal{L}_{scc}$ is shown in the Equation (3).

$$\mathcal{L}_{scc} = \frac{1}{C} \| M - f^{-1}(\hat{M}) \|_1 \tag{3}$$

where $\hat{M}$ represents the seed SC-CAM generated by the data augmentation view, and the SCC loss is averaged over C foreground classes, so it is called a class-specific loss. f^{-1} is the inverse transformation, which is used to restore the spatial order of markers after the image undergoes enhancement, such as flipping. See Section 2.3 for a detailed explanation. This loss is a regularization of the semantic similarity of each branch. $\mathcal{L}_{scc}$ not only regularizes the original image branch but also forces it to focus on the foreground class area to reduce background noise, converting the discriminative area in the foreground and the easily ignored area closer.

In addition, we found in the experiment that after adding a class-specific consistency loss. However, we reduced the false positive pixels. However, the foreground activation of the initial seed was still not high, and we cannot use it directly as a pseudo mask for fully supervised semantic segmentation training. In order to improve the foreground category activation response of the pseudo mask, as shown in Figure 2, we utilize all output tokens to generate a coarse category localization map. Details are as follows:

After L Transformer encoding layer, class tokens and patch tokens are arranged into the final output token $T_{final} \in \mathbb{R}^{(C+N^2) \times D}$, the final class tokens $T_{final-cls} \in \mathbb{R}^{C \times D}$ and the final patch tokens $T_{final-pat} \in \mathbb{R}^{N^2 \times D}$ are defined. $T_{final-cls}$ can obtain the classification score through global average pooling, as shown in Equation (4):

$$\hat{y}(c) = \frac{1}{D} \sum_{i}^{D} T_{\text{final}-cls}(c, i) \tag{4}$$

where $\hat{y}$ is the class prediction, and $c \in 1, 2, ..., C$, $T_{\text{final}-cls}(c, i)$ represent the i^{th} feature of the

ISSN 2616-5775 Vol. 7, Issue 3: 21-31, DOI: 10.25236/AJCIS.2024.070303

c^{th} class token. Finally, we calculate the multi-label soft margin loss $\mathcal{L}_{cls1}$ between the classification score $\hat{y}$ and the image-level ground-truth label y through Equation (5). This enables each class token to capture class-specific information. Similarly, we can obtain different class label classification losses $\mathcal{L}_{cls2}$ in another branch; refer to Equation (5).

$$\mathcal{L}_{cls1} = \frac{1}{C}\sum_{i=1}^{C} y_i \log \sigma(\hat{y}_i) + (1-y_i)\log(1-\sigma(\hat{y}_i)) \tag{5}$$

where σ represents the sigmoid activation function.

$T_{final-pat}$ can be reshaped from a 2-D tensor to a 3-D tensor $T_{final-pat} \in \mathbb{R}^{N\times N\times D}$. At this time, we regard $T_{final-pat}$ as a high-dimensional semantic feature map. Drawing on traditional semantic segmentation tasks, by concatenating a Seg head, we can now obtain a coarse class location map (Coarse-CAM), which is used as the initial seed supplementary; the equation expression is shown in Equation (6). Coarse-CAM is finally converted into class prediction through a global average pooling (GAP) layer, as shown in Equation (7). Similarly, we calculate the multi-label soft margin loss $\mathcal{L}_{coarse-cam1}$ between the classification score $\hat{y}_{coarse-cam}$ and the image-level ground-truth label y through Equation (5). By constraining the classification loss, we can effectively improve the activation response of the Coarse-CAM foreground class. Similarly, we can get different Coarse-CAM and its classification loss $\mathcal{L}_{coarse-cam2}$ in another branch. Since we use horizontal flip as a data augmentation method, all matrices generated by this branch always consistently maintain the shape of the matrix generated by the original image branch.

$$Coarse-CAM = Seg\_head(Resize(T_{final}[C+1:C+N^2])) \tag{6}$$

$$\hat{y}_{coarse-cam} = GAP(Coarse\_CAM) \tag{7}$$

Finally, by element-by-element multiplication, we fuse the specific category initial seed SC-CAM we generated and the rough category location map Coarse-CAM to obtain the Pseudo Mask $M_{Pseudo} \in \mathbb{R}^{C\times N\times N}$, refer to Equation (8). Pseudo Mask can generate pseudo ground-truth (PGT) labels for fully supervised semantic segmentation tasks through some standard refinement methods.

$$M_{Pseudo} = (SC\_CAM) \otimes (Coarse\_CAM) \tag{8}$$

Where $\otimes$ denotes matrix space multiplication and the corresponding position elements are multiplied.

The bi-branch seed generation and class-specific consistency loss proposed in this paper improve the quality of the pseudo mask, allowing it to focus on the foreground object area in the image and reduce the impact of background noise. It also proves the effectiveness of the dual-branch Transformer network architecture on weakly supervised semantic segmentation tasks. Moreover, we do not need complicated post-processing steps like convolutional neural networks. The twin network's initial seeds and rough category positioning maps can be directly generated during the model training process. The training efficiency is also better than that of traditional convolutional neural networks. This can effectively alleviate the pressure of the traditional multi-stage weakly supervised semantic segmentation model.

2.3. Pair-wise Affinity Consistency Loss

Affinity[1, 10] typically refers to the similarity or correlation between different pixels in an image. This similarity or correlation helps the model understand the connections between different regions in the image, thus refining the initial seeds more effectively. In traditional convolutional neural networks, pair-wise affinities are often utilized to refine pseudo-masks and enhance their quality. However, they usually require training an additional affinity network to learn the affinity graph, which significantly burdens model training. In contrast, in previous work[8, 11], it was found that pairwise affinity maps can be directly extracted from the self-attention maps generated by Transformer networks.

ISSN 2616-5775 Vol. 7, Issue 3: 21-31, DOI: 10.25236/AJCIS.2024.070303

In the proposed framework based on the Transformer bi-branch network architecture in this paper, after passing through L Transformer encoder layers, we ultimately obtain attention maps $A_{att} \in \mathbb{R}^{(C+N^2)\times(C+N^2)}$ between all pairs of tokens in the original image and attention maps $A'_{att} \in \mathbb{R}^{(C+N^2)\times(C+N^2)}$ between all pairs of tokens in the augmented view. The affinity matrix refers to the correlation between different pixels. Thus, specific category information is not required. We can extract the attention map between patch tokens, i.e., the affinity matrix $A_{aff} \in \mathbb{R}^{N^2\times N^2}$, from the attention maps between all pairs of tokens. This extraction does not require additional computation or supervision, as shown in Equation (9). The extracted affinity matrix A_{aff} is used to further refine the class-specific pseudo mask M_{Pseudo}, obtaining the pseudo-ground-truth labels $PGT \in \mathbb{R}^{C\times N\times N}$, as shown in Equation (10):

$$A_{aff} = A_{att}[C : C+N^2, C : C+N^2] \tag{9}$$

$$PGT = A_{aff} \bullet M_{Pseudo} \tag{10}$$

Similarly, we designate the affinity matrix generated by the original view branch as A_{aff} and the matrix generated by the augmented view branch as A'_{aff}. We encourage the pairwise relationships between image regions to remain unchanged after transformation. Therefore, we propose the pairwise affinity consistency loss $\mathcal{L}_{pac}$, with the calculation shown in Equation (11):

$$\mathcal{L}_{pac} = \| A_{aff} - f^{-1}(A'_{aff}) \|_1 \tag{11}$$

It is worth noting that when using image augmentation, the appearance of the image and the relative positions of patch annotations are altered. This introduces another issue where the two views' initial seeds (M and $\hat{M}$) and affinity matrices (A_{aff} and A'_{aff}) may not be spatially equivalent, affecting our direct computation of their relative distances. To address this, we employ a simple inverse transformation to reverse the relative positions of annotations and restore their spatial order. At this point, A_{aff} and A'_{aff} have the same spatial order but different values, which may affect the accuracy of our results. This is denoted as f^{-1} in Equations (3) and (11).

Thus, the overall loss function of our approach consists of the bi-branch class tokens classification loss, the Coarse-CAM classification loss, the class-specific consistency loss $\mathcal{L}_{scc}$, and the pairwise affinity consistency loss $\mathcal{L}_{pac}$, as shown in Equation (12):

$$\mathcal{L}_{total} = \frac{1}{2}(\mathcal{L}_{cls1} + \mathcal{L}_{cls2}) + \frac{1}{2}(\mathcal{L}_{coarse-cam1} + \mathcal{L}_{coarse-cam2}) + \mathcal{L}_{scc} + \alpha\mathcal{L}_{pac} \tag{12}$$

where α is a hyperparameter.

3. Experiment and Analysis of Results

3.1. Experimental datasets and evaluation metrics

Public datasets for training, validation, and testing purposes are commonly employed in weakly supervised semantic segmentation. This practice ensures a balanced comparison with preceding methodologies. The ensuing section primarily delineates the task of image-level weakly supervised semantic segmentation; some commonly used representative data sets: SBD[12], PASCAL VOC 2012[13], MS COCO 2014[14], and ImageNet[15]. The approach detailed in this article is primarily trained and evaluated using two datasets, namely PASCAL VOC 2012 and MS COCO 2014. This choice facilitates a comprehensive comparison with other state-of-the-art methodologies.

ISSN 2616-5775 Vol. 7, Issue 3: 21-31, DOI: 10.25236/AJCIS.2024.070303

SBD, a semantic boundary dataset, is utilized for assessing the prediction of semantic contours or boundaries, rather than semantic segmentation directly. However, it can also serve as valuable data augmentation for the semantic segmentation task.

The PASCAL VOC 2012 dataset comprises pixel-level annotations across 21 categories, including background, with 4,369 images drawn from real-world scenes. Among these, 1,464 images are allocated for training, 1,449 for validation, and 1,456 for testing, thus providing a comprehensive resource for research and evaluation. As per the specifications outlined in DeepLab, the SBD dataset serves as supplementary data augmentation for PASCAL VOC 2012, contributing to the enhanced robustness of the model.

MS COCO 2014 is a widely used and expansive dataset primarily focused on scene understanding, encompassing various downstream computer vision tasks such as object detection, instance segmentation, and image captioning. Notably challenging for weakly supervised semantic segmentation research, it offers a more intricate landscape than PASCAL VOC 2012. With 80 distinct foreground object classes and 82,081 training images along with 40,137 validation images sourced from diverse real-world scenarios, MS COCO 2014 encapsulates complex contextual interactions essential for advancing semantic segmentation algorithms.

$$mIoU = \frac{1}{K+1}\sum_{i=0}^{K}\frac{p_{ii}}{\sum_{j=0}^{K} p_{ij} + \sum_{i=0}^{K} p_{ji} - p_{ii}} \tag{13}$$

In image-level weakly supervised semantic segmentation tasks, the commonly used evaluation index is the mean Intersection over Union (mIoU). The calculation method is shown in Equation (13). The size of model parameters: Params (M). They are used to measure segmentation accuracy and model parameters, respectively.

3.2. Experimental Environment

The experiments were conducted on a Ubuntu 20.04 operating system, an NVIDIA A100 Tensor Core GPU, and a device configured with the Pytorch deep learning framework. During the training process, the batch size was set to 64, and the epoch was chosen to be 60. The gradient descent optimization algorithm used was AdamW. For the fully supervised semantic segmentation phase, we adhere to the methodology outlined in previous research[16], employing the DeepLab V1 network based on ResNet38. During testing, we employ multi-scale testing and Conditional Random Fields (CRF)[17] for post-processing to enhance segmentation accuracy.

3.3. Ablation Studies

This section mainly conducts experimental analysis on the PASCAL VOC 2012 dataset and proves the effectiveness of each component on the network. In addition, through ablation experiments, the feasibility of the $\mathcal{L}_{scc}$ and $\mathcal{L}_{pac}$ two losses of and is proved.

(1) Effectiveness of consistency regularization loss:

CAM and Coarse-CAM. In the table, w/o affinity represents the mIoU (%) of the pseudo mask that is not refined using the pairwise affinity matrix (A_{aff}) in the pseudo mask generation stage, and w/o affinity represents the mIoU (%) of the pseudo mask that is refined using the pairwise affinity matrix. We eliminate the two consistency regularization losses in Table 4-1, $\mathcal{L}_{scc}$, which means adding class-specific consistency regularization to the backbone network, and $\mathcal{L}_{pac}$, which means adding pairwise affinity consistency regularization to the backbone network. For more intuitive observation, this chapter adds a pseudo-mask visual comparison of different components on the VOC 2012 data set, as shown in Figure 3. Figure 3 also shows that the method proposed in this article generates a completer and more accurate CAM (pseudo mask) than the CNN method, solving the problem that the initial seeds generated by the convolutional neural network only focus on the discriminative areas in the image.

ISSN 2616-5775 Vol. 7, Issue 3: 21-31, DOI: 10.25236/AJCIS.2024.070303

Table 1: Performance comparison of different proposed components on the PASCAL VOC training set.

baseline*	$\mathcal{L}_{scc}$	$\mathcal{L}_{pac}$	w/o affinity	w/ affinity
√			56.9	60.6
√	√		$59.2_{+2.3}$	$61.8_{+1.2}$
√		√	$59.3_{+2.4}$	$62.6_{+2.0}$
√	√	√	$61.5_{+4.6}$	$63.2_{+2.8}$

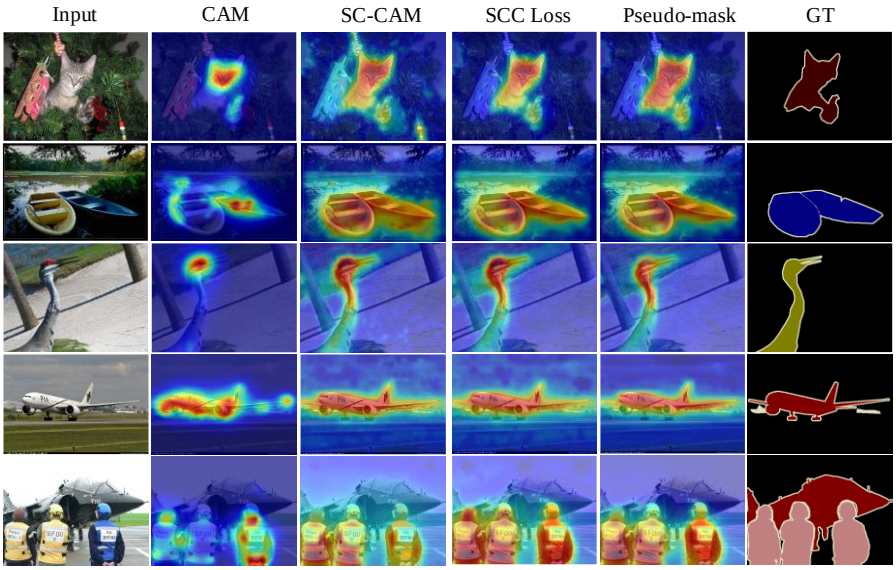

Figure 3: Visualization examples of different object localization maps from different methods

In Table 1, the baseline*[8] reveals our initial challenge in reproducing the segmentation performance of the pseudo mask using the author's provided source code. Upon retraining the benchmark results on our server without pairwise affinity matrix refinement, we achieved a mIoU of 56.9 with affinity refinement and 60.6% without. Introducing a class-specific consistency loss $\mathcal{L}_{scc}$ yielded a notable improvement, raising the mIoU on the PASCAL VOC training set from 60.6% to 61.8%, marking a 1.2% increase. This enhancement underscores the value of leveraging additional supervision information in the enhanced branch of the dual-branch Transformer network. The model more effectively distinguishes foreground from easily ignored areas by enforcing attention on foreground category areas and reducing background noise. Solely incorporating pairwise affinity consistency loss $\mathcal{L}_{scc}$ boosted pseudo mask segmentation accuracy to 62.6%, a 2.0% increase. This demonstrates the efficacy of aligning pairwise affinities between original image branches and enhanced branches, preserving semantic content post-transformation and enhancing the quality of the extracted pairwise affinity matrix from Multi-Head Self-Attention (MHSA). Finally, the simultaneous integration of two regularization losses, $\mathcal{L}_{scc}$, and $\mathcal{L}_{scc}$, resulted in a further improvement, with pseudo-mask segmentation accuracy increasing by 2.8%. Notably, in the absence of pairwise affinity refinement, the mIoU surged by 4.4%, reaching 63.2%, showcasing the superior performance of our proposed regularization method.

(2) Analysis of the impact of different image enhancement methods on pseudo masks

Table 2: Effects of different image enhancement methods on pseudo mask quality.

Image enhancement methods	mIoU (%)
No enhancement	60.6
Rotation	61.9
Vertical flip	60.5
Horizontal flip	**63.2**
Vertical + Horizontal flip	61.9

In the Bi-Trans method proposed in this article, we employed several image enhancement techniques to derive the second view, as detailed in Tables 2, along with their corresponding performance metrics. We refrained from utilizing the scaling enhancement method commonly employed in convolutional networks. This decision was made to facilitate consistency regularization between the original branch

ISSN 2616-5775 Vol. 7, Issue 3: 21-31, DOI: 10.25236/AJCIS.2024.070303

and the enhanced branch. Our experimentation encompassed four image enhancement methods: rotation, vertical inversion, horizontal inversion, and vertical plus horizontal inversion. Remarkably, each image enhancement method improved the network, thus validating Bi-Trans' efficacy. Ultimately, pseudo masks have the most favorable results when solely employing horizontal inversion image enhancement.

3.4. Comparison with State-of-the-art

(1) PASCAL VOC 2012 dataset:

Initially, Table 3 presents the quantitative outcomes of CAM (pseudo mask) and pseudo ground truth (PGT) on VOC 2012. PGT is typically acquired through supplementary post-processing refinement techniques, including widely-used methods such as dense CRF [], PSA [] refinement, and IRN [] refinement. This study adheres to the baseline approach and adopts PSA as the post-processing refinement method. It is noteworthy that its enhancement is not as substantial as that achieved by IRN and online retraining methods.

Table 3: Performance (mIoU) comparison with other methods of pseudo-mask and pseudo-GT

Method	Pub.	Mask	PGT
PSA[1]	CVPR'18	48.0	61.0
SEAM[9]	CVPR'20	55.4	63.6
CONTA[18]	NeurIPS'20	56.2	67.9
AdvCAM[19]	CVPR'21	55.6	68.0
CDA[20]	CVPR'21	55.4	63.4
ViT-PCM[21]	ECCV'22	**63.6**	67.1
SIPE[22]	CVPR'22	58.2	64.7
Ours	-	63.4	**70.6**

Subsequently, Table 4-5 presents the quantitative outcomes of the final semantic segmentation produced by various WSSS methods on VOC 2012. In this context, "Backbone" denotes the backbone model of the segmentation network, while "Sup." indicates the model supervision method, with 'I' representing image-level supervision and 'S' indicating the incorporation of existing saliency maps. For the sake of a fair comparison, we employ the ResNet network as the backbone of the segmentation network in the ultimate segmentation stage. Notably, our method demonstrates competitive results, providing further evidence of the superiority of the Bi-Trans algorithm.

Table 4: Comparison of final segmentation results with other methods on the VOC 2012 validation and test data sets

Method	Bckbone	Sup.	val	test
EPS[23]	ResNet101	I+S	71.0	71.8
L2G[24]	ResNet101	I+S	72.1	71.7
SEAM[9]	ResNet38	I	55.4	63.6
CDA[20]	ResNet38	I	66.1	66.8
ReCAM[25]	ResNet101	I	68.5	68.4
SIPE[23]	ResNet101	I	68.8	69.7
MCTformer[8]	ResNet38	I	70.6	70.3
ToCo[26]	ViT-B	I	69.8	70.5
Ours	ResNet38	I	70.5	70.1

(2) MS COCO 2014 dataset:

Table 5: Comparison of final segmentation results with other methods on the VOC 2012 validation and test data sets

Method	Bckbone	Sup.	val
EPS[23]	ResNet101	I+S	35.7
L2G[24]	ResNet101	I+S	44.2
RCA[27]	ResNet101	I+S	36.8
SEAM[9]	ResNet38	I	31.9
RIB[28]	ResNet101	I	43.8
ReCAM[25]	ResNet101	I	**45.0**
SIPE[23]	ResNet101	I	43.6
MCTformer[8]	ResNet38	I	42.0
ToCo[26]	ViT-B	I	41.3
Ours	ResNet38	I	42.6

ISSN 2616-5775 Vol. 7, Issue 3: 21-31, DOI: 10.25236/AJCIS.2024.070303

Table 5 showcases the quantitative outcomes of the final semantic segmentation attained by various WSSS methods on COCO 2014. This evaluation aims to verify the algorithm proposed in this article for its generalization and performance in complex scenarios. Here, "Backbone" denotes the backbone model of the segmentation network, while "Sup." indicates the model supervision method, with 'I' representing image-level supervision and 'S' indicating the use of existing saliency maps. Remarkably, our method achieves competitive results, surpassing the performance of most algorithms. This outcome underscores the capability of the Bi-Trans method to deliver enhanced segmentation accuracy even in complex scenes.

4. Conclusions

In this paper, we address the issue of the initial seed CAM generated by traditional CNNs, which tend to focus solely on the target discriminative area. Leveraging the robust global modeling capabilities of the Vision Transformer (ViT) model, which holds promising segmentation potential, we introduce a straightforward yet potent network framework termed Bi-Trans. Bi-Trans enriches the network with learnable self-supervised information through an image enhancement branch, thereby mitigating the impact of background noise in the original image. We propose a class-specific consistency loss mechanism that narrows the gap between initial seeds from different perspectives, bolstering their coherence and compelling the original image seeds to exhibit heightened activation in foreground regions. Furthermore, the pairwise affinity consistency loss enhances coherence between affinity matrices derived from diverse perspectives, diminishing the influence of background noise in pseudo masks and elevating their quality. We evaluate Bi-Trans against other state-of-the-art methodologies on two widely used datasets for this task, showcasing the efficacy of our approach. Our work advances the utility of ViT networks in image-level weakly supervised semantic segmentation tasks.

References

[1] Ahn J, Kwak S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4981-4990.

[2] Bearman A, Russakovsky O, Ferrari V, et al. What's the point: Semantic segmentation with point supervision[C]//European conference on computer vision. Cham: Springer International Publishing, 2016: 549-565.

[3] Zhang B, Xiao J, Zhao Y. Dynamic feature regularized loss for weakly supervised semantic segmentation [J]. arXiv preprint arXiv:2108.01296, 2021.

[4] Lee J, Yi J, Shin C, et al. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 2643-2652.

[5] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2921-2929.

[6] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv:2010.11929, 2020.

[7] Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers[C]// Proceedings of the IEEE/CVF international conference on computer vision. 2021: 9650-9660.

[8] Xu L, Ouyang W, Bennamoun M, et al. Multi-class token transformer for weakly supervised semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4310-4319.

[9] Wang Y, Zhang J, Kan M, et al. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 12275-12284.

[10] Ahn J, Cho S, Kwak S. Weakly supervised learning of instance segmentation with inter-pixel relations[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 2209-2218.

[11] Ru L, Zhan Y, Yu B, et al. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 16846-16855.

[12] Hariharan B, Arbeláez P, Bourdev L, et al. Semantic contours from inverse detectors[C]//2011 international conference on computer vision. IEEE, 2011: 991-998.

[13] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge [J]. International journal of computer vision, 2010, 88: 303-338.

ISSN 2616-5775 Vol. 7, Issue 3: 21-31, DOI: 10.25236/AJCIS.2024.070303

[14] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.
[15] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009: 248-255.
[16] Wu Z, Shen C, Van Den Hengel A. Wider or deeper: Revisiting the resnet model for visual recognition [J]. Pattern Recognition, 2019, 90: 119-133.
[17] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv preprint arXiv:1412.7062, 2014.
[18] Zhang D, Zhang H, Tang J, et al. Causal intervention for weakly-supervised semantic segmentation[J]. Advances in Neural Information Processing Systems, 2020, 33: 655-666.
[19] Lee J, Kim E, Yoon S. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4071-4080.
[20] Su Y, Sun R, Lin G, et al. Context decoupling augmentation for weakly supervised semantic segmentation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 7004-7014.
[21] Rossetti S, Zappia D, Sanzari M, et al. Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 446-463.
[22] Chen Q, Yang L, Lai J H, et al. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4288-4298.
[23] Lee S, Lee M, Lee J, et al. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 5495-5505.
[24] Jiang P T, Yang Y, Hou Q, et al. L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 16886-16896.
[25] Chen Z, Wang T, Wu X, et al. Class re-activation maps for weakly-supervised semantic segmentation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 969-978.
[26] Ru L, Zheng H, Zhan Y, et al. Token contrast for weakly-supervised semantic segmentation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 3093-3102.
[27] Zhou T, Zhang M, Zhao F, et al. Regional semantic contrast and aggregation for weakly supervised semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4299-4309.
[28] Lee J, Choi J, Mok J, et al. Reducing information bottleneck for weakly supervised semantic segmentation [J]. Advances in Neural Information Processing Systems, 2021, 34: 27408-27421.