Enhanced Pseudo-Label Generation With Self-Supervised Training for Weakly-Supervised Semantic Segmentation

Zhen Qin, *Member, IEEE*, Yujie Chen, Guosong Zhu, Erqiang Zhou, Yingjie Zhou, *Member, IEEE*, Yicong Zhou, *Senior Member, IEEE*, and Ce Zhu, *Fellow, IEEE*

***Abstract*— Due to the high cost of pixel-level labels required for fully-supervised semantic segmentation, weakly-supervised segmentation has emerged as a more viable option recently. Existing weakly-supervised methods tried to generate pseudo-labels without pixel-level labels for semantic segmentation, but a common problem is that the generated pseudo-labels contain insufficient semantic information, resulting in poor accuracy. To address this challenge, a novel method is proposed, which generates class activation/attention maps (CAMs) containing sufficient semantic information as pseudo-labels for the semantic segmentation training without pixel-level labels. In this method, the attention-transfer module is designed to preserve salient regions on CAMs while avoiding the suppression of inconspicuous regions of the targets, which results in the generation of pseudo-labels with sufficient semantic information. A pixel relevance focused-unfocused module has also been developed for better integrating contextual information, with both attention mechanisms employed to extract focused relevant pixels and multi-scale atrous convolution employed to expand receptive field for establishing distant pixel connections. The proposed method has been experimentally demonstrated to achieve competitive performance in weakly-supervised segmentation, and even outperforms many saliency-joined methods.**

Manuscript received 28 June 2023; revised 2 December 2023; accepted 6 February 2024. Date of publication 9 February 2024; date of current version 12 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62072074, Grant 62076054, Grant 62027827, Grant 62002047, Grant 62171302, and Grant 62020106011; in part by the Sichuan Science and Technology Support Plan under Grant 2022YFQ0045, Grant 2022YFS0220, Grant 2021YFG0131, Grant 2023YFS0020, Grant 2023YFS0197, and Grant 2023YFG0148; in part by the CCF-Baidu Open Fund under Grant 202312; in part by the 111 Project under Grant B21044; in part by the Sichuan Science and Technology Program under Grant 2023NSFSC1965; and in part by the Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China under Grant ZYGX2021YGLH212 and Grant ZYGX2022YGRH012. This article was recommended by Associate Editor L. Chiariglione. *(Corresponding author: Yingjie Zhou.)*

Zhen Qin, Yujie Chen, Guosong Zhu, and Erqiang Zhou are with the Network and Data Security Key Laboratory of Sichuan Province, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: qinzhen@uestc.edu.cn; 202021090126@std.uestc.edu.cn; 202111090804@std.uestc.edu.cn; zhoueq@uestc.edu.cn).

Yingjie Zhou is with the College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: yjzhou@scu.edu.cn).

Yicong Zhou is with the Department of Computer and Information Science, University of Macau, Macau, China (e-mail: yicongzhou@um.edu.mo).

Ce Zhu is with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: eczhu@uestc.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2024.3364764.

Digital Object Identifier 10.1109/TCSVT.2024.3364764

***Index Terms*— Semantic segmentation, weakly-supervised learning, attention transfer mechanism, class attention/activation maps.**

I. Introduction

Semantic segmentation aims to predict object class labels and pixel-specific object masks, which can locate different types of objects that exist in various images. It has become one of the most important, complex and challenging fields in computer vision research. With the advancement of Convolutional Neural Networks (CNNs), numerous fully-supervised semantic segmentation models have been proposed in recent years, which makes the segmentation accuracy be greatly improved. Due to the data-hungry nature of deep CNN, segmentation requires a large number of training images with ground truth labels, which are typically manually annotated. This is particularly true for segmentation tasks, as they necessitate time-consuming and costly pixel-level labeling for effective training. This leads to constraints on the availability of annotated data for existing models.

Weakly supervised learning is an approach to address these limitations, which employs supervision in the form of image-level labels that are less precise but more economical than semantic segmentation masks [1]. In recent years, various forms of image-level labels have been investigated for the purpose of weakly-supervised semantic segmentation, including image-level labels, point labels, scribbles and bounding boxes. These works aim to achieve fully supervised segmentation while only employing weakly semantic segmentation with image-level classification labels, which could be obtained at a low annotation cost. To achieve this goal, high-quality pseudo-labels are generated by neural networks as a substitute for manual labels in our work. Then the generated pseudo-labels are employed to train the semantic segmentation network.

Most of weakly-supervised segmentation methods attempt to extract semantic information from image-level labels [2]. The mainstream of them rely on class attention/activation maps (CAMs) [3], as the representation of locating targets. CAMs are usually generated based on the classification network, where the classifier focuses on salient regions of recognizable objects and aims to effectively classify them into different categories. Riding on the success of these advanced methods, our approach also tries to generate CAMs as pseudo-labels for the training of semantic segmentation. The boundaries and details of targets in existing weakly-supervised segmentation methods are usually not accurate enough, which limits the performance of the trained semantic segmentation network. Although the leveraging of CAMs significantly improves the performance of weakly-supervised segmentation, there is still a large gap with the fully-supervised segmentation method which uses accurate pixel-level labels.

To further improve the performance of weakly-supervised segmentation, a novel method is proposed by generating the CAMs as pseudo-labels for the training of semantic segmentation, which could activate the regions of recognizable targets with accurate boundaries. The recognizable targets correspond to image-level classification labels, and corresponding pixels can provide higher activation scores for the classification network compared with those for background pixels. The pixels with respect to recognizable targets, suppressing the background pixels, could be regarded as pseudo-labels. The modules, i.e., the attention transfer module and the pixel relevance focused-unfocused module, are designed to improve the quality of pseudo-label generation in our method.

Inspired by contrastive learning, we propose the attention transfer module to generate better CAMs, which effectively extracts and activates hidden pixels of targets, transferring salience attention to overall target attention in a self-supervised manner. We first apply image augmentation techniques to adjust the contrast and brightness of the original image before inputting it into the classification network. Then we merge the output CAMs with the original image CAM to transfer classifier attention. The reason why we adjust only the contrast and brightness is that this strategy can optimize the quality of generated CAMs in certain aspects without changing the semantic content of the images. Enhancing contrast and brightness makes the important features of objects in the generated CAMs more salient, leading to accurate localization of the target. Decreasing contrast and brightness enriches the semantic of targets in the generated CAMs, making the semantic information more comprehensive. By merging enhanced and weakened CAMs, the network can be trained to simultaneously learn more accurate target localization and more complete semantic information. Finally, we define a similarity loss function to optimize the classification network, making the final used CAM closer to ground truth. The proposed attention transfer module enables the classifier to not only attend to salient object regions, but also capture suppressed activation regions which belongs to targets.

Furthermore, to enhance the framework's capacity of extracting features while avoiding the loss of details caused by convolution operations, an effective pixel relevance focused-unfocused module has also been designed to generate higher-quality CAMs. For the pixel relevance focused component in this module, channel attention and spatial attention mechanisms are employed to extract pixel relationships and acquire finer pixel-wise information. In this way, the framework is capable of accurately focusing on important features and achieving accurate locating of objects. For the unfocused component of pixel relevance, an improved 4-layer multi-scale atrous convolution is employed to expand the receptive field, while preventing the excessive loss of details. Unlike previous atrous convolutions, the improved 4-layer multi-scale atrous convolution employed in the framework has different dilation factors for each layer, allowing for better extraction of global information. Context association is also employed to build connections between pixels, facilitating a better understanding of features.

Our method overcomes the limitations that existing CAMs only focuses on salient object regions, while generating pseudo-labels which represents complete semantic information with accurate boundaries of targets. The generated pseudo-labels can be used in training more accurate semantic segmentation models, further narrowing the gap between semantic segmentation in weakly-supervised manner with image-level labels and fully-supervised manner with pixel-level labels. The contributions are summarized as follows:

- The novel framework employing an effective self-supervised training strategy is proposed to generate pseudo-labels for semantic segmentation. The framework learns accurate localization of important object features and complete semantic information while preserving the original semantics, which is capable of generating high-quality class attention maps as pseudo-labels.
- An effective pixel relevance focused-unfocused module is designed in the framework, which adopts channel and spatial attention mechanisms to focus on important features, alone with the improved 4-layer multi-scale and context association to expand the receptive field. This module significantly improves the capacity of framework in deeply extracting semantic information, while avoiding excessive loss of details.
- The proposed method has been experimentally demonstrated to achieve competitive performance in weakly-supervised segmentation on the PASCAL VOC 2012 dataset [4], and even outperforms many saliency-joined methods.

II. Related Work

This section reviews semantic segmentation models closely related to our method. We first introduce the existing weakly supervised approaches for the task, and then discuss self-supervised learning works. Finally, we discuss the advantage and drawbacks of traditional atrous convolution.

A. Weakly Supervised Semantic Segmentation

Unlike fully-supervised semantic segmentation, which requires pixel-wise labels for images, weakly-supervised semantic segmentation methods employ low-cost labeling, such as bounding boxes, scribbles, and image-level classification labels. Weakly-supervised semantic segmentation has been significantly boosted by employing CAMs to generate pseudo labels [5]. Considering that the bounding boxes generated by classification network contain abundant semantic and objective information [6], most existing weakly-supervised semantic segmentation methods refine the CAMs generated by the image classifier to approximate the segmentation mask. Proposed by Ahn et al. [7], AffinityNet trains an additional network to learn similarities between the pixels, which often generates a transition mix and multiplies with CAM to adjust its activation coverage. Also proposed by Ahn et al. [8], IRNet generates a transition matrix from the boundary activation map and extends the method to achieve weakly supervised instance segmentation and weakly-supervised semantic segmentation. Hao et al. [9] proposed a novel network that establishes context correlation through CAMs to achieve efficient video segmentation. Chen et al. [10] proposed an effective method of reactivating the converged CAM with binary cross-entropy loss by using softmax cross-entropy loss, improving the quality of CAMs generation while reducing computational overhead. Proposed by Wang et al. [11], SEAM aims to refine CAMs using a pixel correlation module that captures context appearance information for each pixel and alters the original CAMs by using learned affinity attention maps. Proposed by Sun et al. [12], SSA directly utilizes semantic structural information to expand CAM during the inference stage, resulting in high-quality CAMs without incurring any additional training cost. Proposed by Li et al. [13], a novel and interesting context-based tandem network is designed for semantic segmentation by effectively exploring the channel context and the spatial context, which significantly improves the segmentation performance. Proposed by Sun et al. [14], a novel strategy that decompose the backbone parameters into three matrices, and one of these matrices is fine-tuned by adjusting its singular values while keeping the other two frozen during training, addressing over-fitting in weakly supervised semantic segmentation.

In general, current weakly supervised semantic segmentation methods still rely on generating CAMs as pseudo-labels to train semantic segmentation networks. The improvements are mostly centered around enhancing the quality of generated CAMs. However, previous methods often struggle to optimize the accuracy of focusing on key features while enriching the completeness of semantic information. To address this issue, we integrate adjusted CAMs to possess both comprehensive semantic information and accurate important feature localization simultaneously.

B. Self-Supervised Learning

Instead of using massive annotated labels to train network, self-supervised learning approaches aim at designing pretext tasks to generate labels without additional manual annotations, which is applied in relative position prediction, spatial transformation prediction, image inpainting [15], image colorization and seismic image analysis [16]. Proposed by Qian et al. [17], FR-Net is designed to separate and remove footprint noise from the image in self-supervised manner, which can be regarded as an important application of self-supervised learning. To some extent, the generative adversarial network can also be regarded as a self-supervised learning approach that the authenticity labels for discriminator do not need to be annotated manually [18]. Labels generated by pretext tasks provide self-supervision for the network to learn a more robust representation [19]. The feature learned by self-supervision can replace the feature pretrained on some tasks, such as detection and part segmentation. Proposed by Liu et al. [20], a novel method for self-supervised time series anomaly detection and clustering is designed. Proposed by Zhou et al. [21], a novel strategy is adopted to transform the input data into meaningful representations, hidden representation, reconstruction residual vector, and reconstruction error that could be used for anomaly detection in self-supervised manner.

Overall, self-supervised methods come in various forms, but their common objective is to generate labels from existing images without relying on manual annotations, serving to economize the cost of manual labeling. In our work, to train our Attention Transfer Module for generating higher-quality CAMs without manual labeling, we adjust the contrast and brightness of the original images to generate corresponding CAMs as supervision. Specifically, CAMs generated after enhancing contrast and brightness exhibit more accurate localization of key features, while those generated after decreasing contrast and brightness contain more comprehensive semantic information. By combining these two types of CAMs, we obtain high-quality CAMs as ground truth, which are used to supervise the training of the CAMs generation network in the Attention Transfer Module.

C. Atrous Convolution In Weakly-Supervised Methods

Many weakly-supervised semantic segmentation methods prefer using atrous convolution in stead of ordinary convolution in the classification network as the backbone in order to capture richer semantic information and achieve better performance. The fully convolutional approach has been proved to be effective for semantic segmentation, but frequent max-pooling and striding may lead to a severe reduction in the spatial resolution of the obtained feature maps. Proposed by Noh et al. [22], deconvolution has been used to restore the spatial resolution, but the holes convolution should be advocated because the receptive field of atrous convolution is larger. When the amount of parameters is certain, ordinary convolution [23] can only extract features of small blocks, while hole convolution can increase the hole rate to make more overlapping sampling areas on the input feature map for each sampling, so as to obtain denser characteristic response. However, atrous convolution also has some drawbacks, such as the inability to extract some key local information and poor performance in object segmentation for small targets. Therefore, we have improved the traditional atrous convolution, as illustrated in III-C. Unlike traditional atrous convolution that use fixed dilation rates, our approach employs a 4-layer multi-scale atrous convolution with varying dilation factors in the convolutional layers. Moreover, the dilation rates are progressively increased with the network depth. This allows our network to significantly expand the receptive field of the convolutional kernels, enabling a better understanding of global information without adding extra parameters or computational load. As a result, this enhances the performance and generalization capabilities of the network.

In summary, atrous convolution is widely employed in weakly supervised methods. Many weakly supervised methods build upon traditional dilated convolutions to further reduce computational cost or increase receptive fields. In our work,

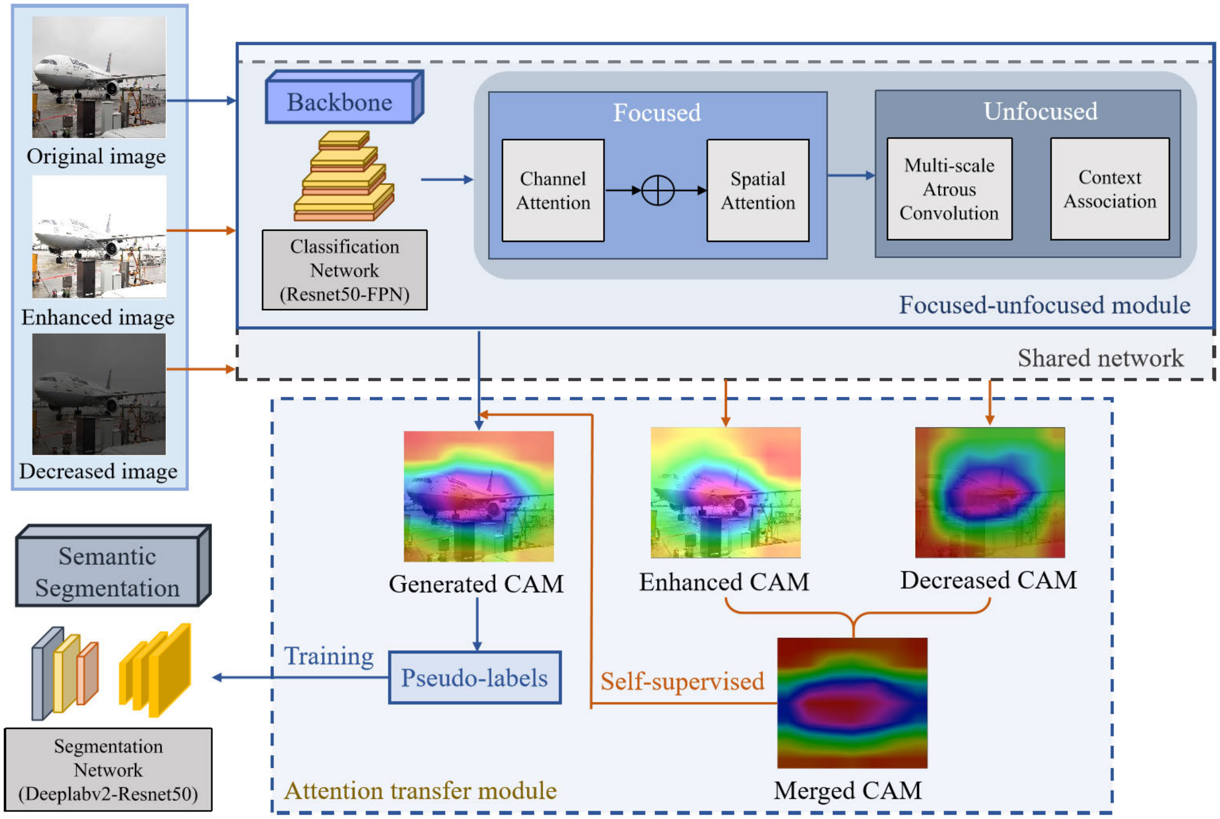


Fig. 1. The overall structure of the proposed method. In this figure, the shared network refers to a network with the same structure as the focused-unfocused module, taking adjusted image as input. In the structure, the backbone is a classification network used to generate CAMs after training. In focused-unfocused module, the focused part is used to help backbone focus on important salient regions, while the unfocused part is used to expand the receptive field of backbone. The attention transfer module adopts data augmentation operation and fuses enhanced and decreased CAMs to obtain the merged CAMs with richer semantics, which are used to retrain the backbone for generating CAMs with more complete semantics as pseudo-labels. The generated CAMs are used to train semantic segmentation networks.

a 4-layer multi-scale atrous convolution is employed, which significantly expands the receptive field of the convolutional kernel without increasing parameters or computational load. This allows for better extracting of global information, leading to improved network performance and generalization capabilities.

In our work, we ameliorate the issues of inaccurate feature localization and incomplete semantic information in the generation of CAMs from previous related work by designing a self-supervised training attention transfer module. This module, in conjunction with the focused-unfocused module that focuses on context and key features, aids in the improved extraction of features.

III. Methodology

This section introduces the proposed method in details. Firstly, we present the preliminary of our work. Then we introduce the proposed attention transfer module. The designed focused-unfocused module is integrated into the network to further improve the consistency of prediction. Finally, the loss design of our network is discussed. The overall structure of our method is shown in Fig. 1.

Channel and spatial attention mechanisms are employed to focus on important features in the Focused-Unfocused module, which also incorporates multi-scale atrous convolution and context association to expand the receptive field for providing inconspicuous semantic information. This assists the Attention Transfer module in generating CAMs with precise feature information, addressing the issues of important features being overlooked and incomplete semantic information of target that exist in previous CAMs generation methods.

A. Preliminary

The conventional CAM of a single image highlights the most representative areas of each class. Therefore, when generating the CAM of the same class for image patches, the model focuses on finding the key features of the class based on partial observation of the object.

For a class c, the CAM is a feature map indicating the discriminative regions of the image which helps the classifier to make the decision that whether the target object belong to class c. As an activation map, the confidence of CAM at location (x, y) is calculated as follows:

$$M_c(x, y) = \sum_k w_k^c f_k(x, y) \tag{1}$$

where $f_k(x, y)$ denotes the activation of a channel k in the last convolutional layer at location (x, y), and w_c denotes the weights from $f_k(x, y)$ via global average pooling by SoftMax function. In the classification task, a higher value of CAM indicates a greater contribution for classification.

However, it is easy to find that the properties of classification and segmentation function are different. The classification

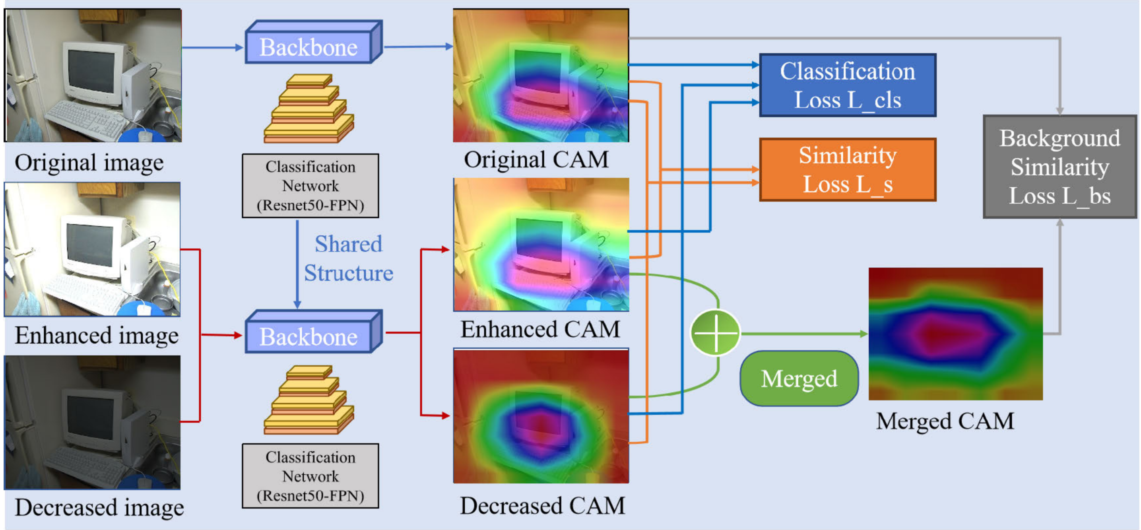


Fig. 2. The structure of the proposed attention transfer module elaborated in III-B and the designed loss functions illustrated in III-D. In this figure, image has been enhanced with contrast and brightness, and enhanced CAM is the CAM generated from the enhanced image. Decreased image has been decreased with contrast and brightness, and decreased CAM is the CAM generated from the decreased image. Merged CAM combines the strengths of both enhanced and decreased CAMs for supervising the generation of CAMs.

task focuses on the salient area of the object, so the CAMs generated by the single classification network will activate the most recognizable area of the object and suppress other areas of the target. This does not meet the goals of the segmentation task. Semantic segmentation focuses on the target as a whole, not just salient regions. Therefore, the attention transfer module is designed to activate other inhibited areas of the target, combining with the focused-unfocused module to extract refined features and expand the receptive field, to obtain optimized CAMs with more semantic information that can be used as pseudo-labels for the training of semantic segmentation.

B. Attention Transfer Module

To allow the CAMs preserving target salient regions while avoiding suppressing semantic segmentation focus regions, the attention transfer module is proposed, which adopts image augmentation strategy to provide richer semantic information for semantic segmentation task in self-supervised manner. The structure of the proposed attention transfer module is shown in Fig. 2.

Since we only used image-level labels, the information obtained from labels is insufficient to support the semantic segmentation task. To solve this, the image augmentation processing is adopted, including two operations: enhancing image contrast and brightness, and decreasing image contrast and brightness. For a single pixel (x, y) in the image, the adjustment of new pixel value $g'(x, y)$ is formulated as follows:

$$g'(x, y) = \alpha g(x, y) + \beta \tag{2}$$

where $g(x, y)$ denotes the original pixel value. α is used to control the intensity of contrast between pixels, and β is used to adjust the brightness of the picture. The visualization of CAMs adjustment is shown in Fig. 3. As is evident from the figure, the enhancement makes the distinction between the instance area more obvious, which enables the network to learn the information of the target area and boundary more clearly. Decrease significantly reduces the recognition degree of significant regions, which also reduces the dependence on significant regions, activating the suppressed regions.

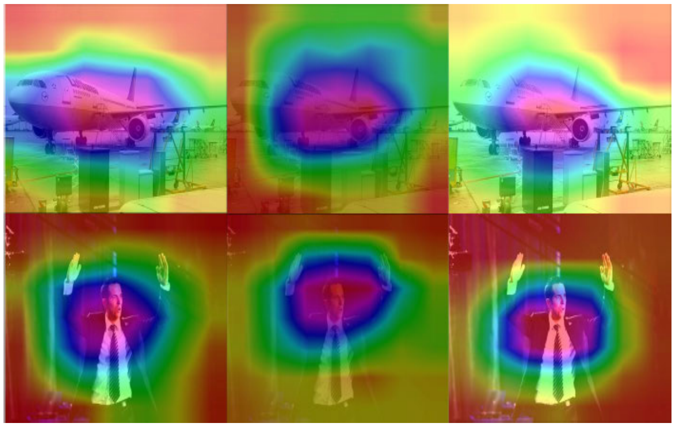


Fig. 3. The visualization of CAMs adjustment. The left column is the CAMs generated from original images, the middle column is generated from images with decrease of contrast and brightness, and the right column is generated from images with enhancement of contrast and brightness.

Based on the aforementioned analysis, the CAM with enhancement has higher activation scores in the most recognizable regions and boundaries of the target, while the CAM with decrease partly activates other regions which are important for semantic segmentation. The combination of these two CAMs outputs the semantic information needed for the semantic segmentation task. For a point x in the merged CAM, the confidence calculation of x belonging to category C can be divided into four cases:

The first case is when point x of both enhanced CAM $M_e(x)$ and decreased CAM $M_d(x)$ belongs to category C, the confidence that corresponding point x of the merged CAM belongs to category C is calculated as $\mathcal{P} = max(\mathcal{P}_e(x), \mathcal{P}_d(x))$. where

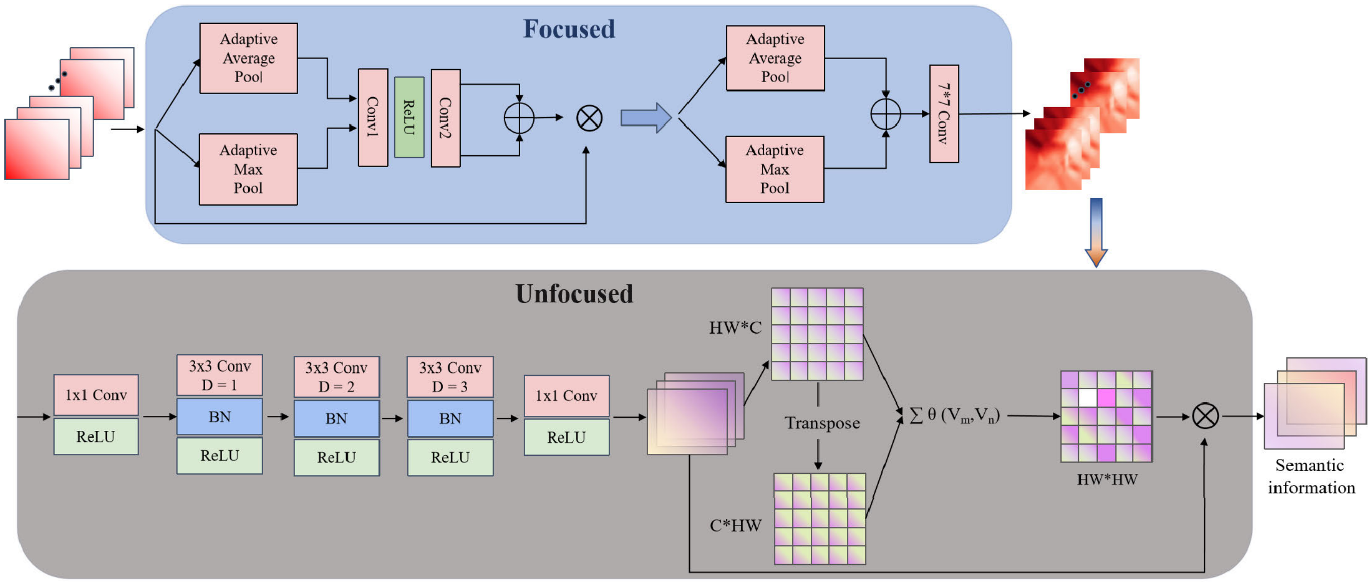


Fig. 4. The structure of the proposed focused-unfocused module. In this module, D in 3×3 $Conv$ denotes dilation factor, and BN denotes batch normalization. The focused part adopts channel and spatial attention mechanisms to focus on important features, and the unfocused part adopts multi-scale atrous convolution and context association to expand the receptive field for providing inconspicuous semantic information.

$\mathcal{P}_e(x)$ and $\mathcal{P}_d(x)$ denote the confidence of x belonging to this category in the adjusted CAM. The second case is when point x of $M_e(x)$ or $M_d(x)$ belongs to category C, while the maximum values of $\mathcal{P}_e(x)$ or $\mathcal{P}_d(x)$ is greater than 0.5, $\mathcal{P} = max(\mathcal{P}_e(x), \mathcal{P}_d(x))$. The third case is when point x of $M_e(x)$ or $M_d(x)$ belongs to category C, but the values of both $\mathcal{P}_e(x)$ and $\mathcal{P}_d(x)$ are smaller than 0.5, $\mathcal{P} = 0$. The fourth case is when point x of neither $M_e(x)$ nor $M_d(x)$ belongs to category C, $\mathcal{P} = 0$. In this way, the merged CAMs can be obtained to supervise the final generated CAMs as pseudo-labels for the training of semantic segmentation in weakly-supervised manner.

In the absence of pixel-level labels as semantic segmentation supervision, the processed CAMs are adopted as supervision to narrow the supervision gap between self-supervision and fully-supervision, obtaining more semantic information.

In the previous operation, we can obtain the CAM M_e generated from image with enhancement, the CAM M_d generated from image with decrease, the CAM M_o generated from original and the CAM $\widehat{M_o}$ combined of M_e and M_d. In feature space, M_e and M_d should be similar to and tend to be consistent with $\widehat{M_o}$, because only the contrast and brightness of the image are adjusted, with the information contained in image unchanged. Therefore, a similarity loss L_s is proposed to better extract the semantic features of images. Furthermore, background similarity between M_o and $\widehat{M_o}$ is also used for self-supervised learning. We activate the background regions of M_o and $\widehat{M_o}$ with background activation scores reassigned. For the foreground, activation scores of the most recognizable pixel are maintained, and the scores of remaining points are set to 0. It is supposed that the background regions and target regions in M_o and $\widehat{M_o}$ are the same. Based on the above design, a background similarity loss L_{bs} is also proposed, with L_s illustrated in subsection III-D.

C. Focused-Unfocused Module

In general, the results of semantic segmentation are positively related to the information provided by labels. However, without pixel-level labeling, image-level classification labeling can provide very limited useful information. Therefore, the focused-unfocused module is designed to further explore the semantic information of pixels, as shown in Fig. 4.

The channel attention mechanism [24] and spatial attention mechanism [25] are introduced in the network to focus on important feature information. Meanwhile, a multi-scale atrous convolution layer [26] is also introduced to expand the receptive field of the remaining regions, i.e., expanding the receptive field of remaining regions to alleviate the influence of insufficient semantic information from its image-level label. However, traditional atrous convolution has defects, such as the lack of partial local information and difficulties in establishing the association of remote features, which is limited in the segmentation of small targets. Therefore, following the hybrid dilated convolution [27], a 4-layer multi-scale atrous convolution with dilation factors of 0, 1, 2, 3 is designed. Each layer is designed with an atrous convolutional followed by an activation function, which activates important regions and suppress irrelevant regions.

To further explore the semantic information for segmentation from images without pixel-level labels, we design a context association submodule, for exploring the context relation of pixels in a deep level with semantic information expanding. The cosine distance is also adopted to measure pixel context similarity. The detailed design is presented as follows. First, the feature map after the designed 4-layer multi-scale atrous convolution is reshaped, with dimension reduced from $H \times W \times C$ to $HW \times C'$. The reshaped feature map F_0 is transposed to get F_0^T, with size of $C' \times HW$. Then the cosine distance between pixels of F_0 and F_0^T is calculated, with activation and normalization operations to suppress irrelevant regions, for obtaining the context relation weight map M_c, with size of $HW \times HW$. The calculation of cosine distance between pixel V_m and pixel V_n is shown in formulation 3, where Φ denotes the reshaping process. Finally, input features are multiplied with M_c to get the weighted feature map, with size of $HW \times C$, followed by reshaped to $H \times W \times C$ as final feature map.

$$\theta(V_m, V_n) = \frac{\Phi(V_m)^T \Phi(V_n)}{|\Phi(V_m)| \times |\Phi(V_n)|} \tag{3}$$

By introducing channel attention mechanism, spatial attention mechanism, 4-layer multi-scale atrous convolution and context association submodule, we explore more semantic information from images without pixel-level labels, optimizing the generated CAMs as illustrated in subsection III-B, for better serving as pseudo-labels to guide the training of semantic segmentation task.

D. Loss Functions

1) Classification Loss: Since the provided labels are image-level for classification, we select the classification network ResNet50-FPN [28] as the backbone in our framework. Accordingly, the cross entropy loss is adopted to train the classification network, and both the original images and the adjusted images are input as training data. The classification loss is formulated as follows:

$$L_{cls} = -\frac{1}{n}\sum_{x}[y \ln a + (1-y)\ln(1-a)] \tag{4}$$

where n denotes the number of samples, and x denotes the prediction vector dimension. y denotes the real value after one-hot coding corresponding to the label on x dimension, with the value of 1 as truth or 0 as false, and a denotes the prediction label, with the range of 0 to 1.

2) Similarity Loss: It is a common truth that adjusting the image contrast and brightness does not change the semantic information of the image. Therefore, when projected into feature space, the CAM M_o generated from original image, the CAM M_e generated from image with enhancement and the CAM M_d from image with decrease should be consistent. We use L1 loss to train the network, making the network extract obvious features in the image, as well as improving the robustness of network in contrast and brightness, as formulated as follows:

$$L_s = \frac{\|M_o - M_d\|_1 + \|M_o - M_e\|_1}{2} \tag{5}$$

3) Background Similarity Loss: CAMs are generated by classification network and iteratively optimized by classification loss of foreground target. However, since the training is based on image-level labels, it is inevitable that the semantic information of many pixels will be ignored, which are learned as background pixels and set as zero vector. In this way, the gradient cannot be generated to learn feature representation through back propagation, leading that a great deal of information provided by background pixels is ignored. Therefore, we design the background similarity loss to explore the semantic information of background. For the CAM M_o generated from original image and the CAM $\widehat{M_o}$ combined of M_e and M_d, as illustrated in subsection III-B, we segment the background for M_o and $\widehat{M_o}$ to get the background activation map, M_b and $\widehat{M_b}$, respectively. L1 loss is also used as the background similarity loss, formulated as follows:

$$L_{bs} = \|M_b - \widehat{M_b}\|_1 \tag{6}$$

By jointly considering classification loss, similarity loss and background similarity loss, the overall loss L for optimizing the architecture weight is formulated as follows:

$$L = L_{cls} + \lambda_s L_s + \lambda_{bs} L_{bs} \tag{7}$$

where λ_s and λ_{bs} denote the hyper-parameters to balance overall loss. L_{cls} is used for roughly locating target regions. L_s is used for providing additional supervision and narrowing the gap with fully-supervision. L_{bs} is used for combining background regions with salient regions to obtain the ignored information of original CAM M_o.

IV. Experiments

In this section, we evaluate the performance of the proposed method for semantic segmentation. We first briefly introduce the dataset used in experiments, and illustrate the implementation details including parameter settings and data preprocessing. We then evaluate the performance of the proposed method and compare with other state-of-the-art methods. Finally, we present ablation studies to demonstrate the contribution of different components in our method.

A. Datasets

Our experiments were conducted on the PASCAL VOC 2012 dataset [4], which contains the images and labels needed for the classification task, detection task, segmentation task, behavior recognition and human layout detection task. This dataset consists of 21 categories, including 20 foreground pixel categories and 1 background pixel category, where images are divided by authority, with 1,464 used for training, 1,449 for verification and 1,456 for testing.

Furthermore, following the general experimental strategy of semantic segmentation, we extract additional labels from SBD dataset [47] and extend the original training set from SBD dataset. According to statistics, 11,355 images in the SDB data set are actually included in the VOC 2012 dataset, but only 1,462 images in the VOC 2012 dataset can be used to train the semantic segmentation task. After introducing the data of SDB dataset, 10,582 images with image-level labels were used for training and 1,449 images for validation in total.

B. Implementation Details

In the image augmentation process as illustrated in subsection III-B, we set $\alpha = 1.5$ and $\beta = 300$ for the formulation 2 to enhance the image contrast and brightness, and we set $\alpha = 0.8$ and $\beta = 150$ to decrease the image contrast and brightness.

Our framework was trained with 10,582 images with image-level classification labels from the designed training-set.

In this process, the poly strategy is adopted, for dynamically adjusting the learning rate, which is formulated as follows:

$$lr = lr_0 \times (1 - \frac{epoch}{max\_epoch})^m \tag{8}$$

where lr_0 denotes the initial learning rate, set as 0.1, and m is the momentum, set as 0.9. *epoch* denotes to the current number of iterations, and *max_epoch* denotes the number of iterations required to complete the training. In this way, the update of the last iteration is taken into account to accelerate the learning and reduce the oscillation phenomenon in the process of model convergence.

To evaluate the effect of the proposed weakly-supervised learning method, we use the pseudo-labels generated by our method to train a fully-supervised semantic segmentation network, where the DeepLab v2-ResNet50 [23] is adopted as the semantic segmentation network to evaluate the quality of generated pseudo-labels. The semantic segmentation accuracy with respect to the model pretrained with generated pseudo-labels can be used to evaluate the performance of our method. In the training of semantic segmentation network, 10,582 images from the designed training-set were also used for training, but the labels were replaced with generated pseudo-labels. The hyper-parameters λ_s and λ_{bs} in the overall loss are set to 0.5 each, in order to balance the overall loss.

C. Benchmarks

Following the commonly used semantic segmentation evaluation scheme [48], the main evaluation metric is Mean Intersection over Union (mIoU), with the tests performed on 20 categories and background of the PASCAL VOC 2012 dataset. We compared multiple state-of-the-art methods for semantic segmentation in weakly-supervised manner, which is performed on PASCAL VOC 2012 dataset, as illustrated in Table I.

In the Table I, saliency denotes the additional supervision of significance regions joined in the weakly-supervised training, which theoretically results in a better performance than methods without saliency. However, our method not only outperformed the compared methods without saliency, but also achieved the competitive performance compared with those of saliency-joined methods.

To demonstrate that our methods achieves competitive performance for objects with different shapes and sizes, we also performed the quantitative comparison with other weakly-supervised semantic segmentation methods on 20 categories and background category, where the test images are from the validation set of Pascal VOC 2012 dataset, as illustrated in Table II.

TABLE I

QUANTITATIVE COMPARISON OF WEAKLY-SUPERVISED SEMANTIC SEGMENTATION METHODS ON PASCAL VOC 2012 DATASET, WHERE MIOU IS ADOPTED AS THE EVALUATION METRIC

Methods	Backbone	Saliency	val ↑	test ↑
AffinityNet (CVPR18) [7]	ResNet38 [29]	✘	61.7	63.7
DSRG (CVPR18) [30]	ResNet101 [29]	✔	61.4	63.2
SeeNet ((NeurIPS18) [31]	ResNet38 [29]	✔	63.1	62.8
IRNet (CVPR19) [8]	ResNet50 [29]	✘	63.5	64.8
SSDD (ICCV19) [32]	ResNet38 [29]	✘	64.9	65.5
CIAN (AAAI20) [33]	ResNet101 [29]	✔	64.3	65.3
Zhang *et al.* (ECCV20) [34]	ResNet50 [29]	✔	66.6	66.7
Sun *et al.* (ECCV20) [35]	ResNet101 [29]	✔	66.2	66.9
SEAM (CVPR20) [11]	ResNet38 [29]	✘	64.5	65.7
Chen *et al.* (ECCV20) [36]	ResNet50 [29]	✘	65.7	66.6
Chang *et al.* (CVPR20) [37]	ResNet38 [29]	✘	66.1	65.9
Wang *et al.* (IJCV20) [38]	ResNet38 [29]	✘	64.3	65.4
Araslanov *et al.* (CVPR20) [39]	ResNet38 [29]	✘	62.7	64.3
ICD (CVPR20) [40]	ResNet101 [29]	✘	64.1	64.3
CONTA (NeurIPS20) [41]	ResNet38 [29]	✘	66.1	66.7
OAA++ (TPAMI21) [42]	ResNet101 [29]	✘	64.9	66.3
OAA++$^+$ (TPAMI21) [42]	VGGNet [43]	✘	63.7	63.2
Su *et al.* (ICCV21) [44]	ResNet38 [29]	✘	66.1	66.8
Zhang *et al.* (ACMMM21) [44]	ResNet38 [29]	✘	63.9	64.8
Ru *et al.* (CVPR22) [45]	MiT-B1 [46]	✘	66.0	66.3
Ours	ResNet50 [29]	✘	**67.3**	**67.5**

As shown in Table II, the performance of our method is superior to those of other compared weakly-supervised semantic segmentation methods on accuracy in many categories of semantic segmentation, where these categories have different shapes and sizes. Due to differences in the size, shape, texture, feature patterns, and surrounding environments of different object categories, the significant difference exists in the segmentation accuracy of these categories. The semantic segmentation results of our method and compared state-of-the-art weakly-supervised methods are also provided, to further illustrate that our method has achieved excellent performance, as shown in Fig. 5.

D. Ablation Studies

In this subsection, we performed ablation studies on the two core modules of our method, i.e., attention transfer module and focused-unfocused module to evaluate the contribution of them in our method.

1) Effect of the Attention Transfer Module: As illustrated in subsection III-B, we adjust the contrast and brightness of the image, obtaining CAM M_e generated from image with enhancement and CAM M_d generated from image with decrease. To evaluate the contribution of introducing adjusted CAMs to the training of our framework for generating final CAMs as pseudo-labels, we introduce M_e separately, M_d separately, and both M_e and M_d to the base model to compare the quality of pseudo labels generated under different conditions, respectively. The quality of generated pseudo-labels is evaluated by comparing with semantic segmentation ground truth labels, as illustrated in Table III.

The evaluation in Table III shows the contribution of introducing M_e and M_d, which obviously improve the quality of pseudo-labels. Note that, the base model of this ablation study was based on the introduction of complete focused-unfocused module, and only used M_o to generate pseudo-labels. Furthermore, the visualization of loss convergence curves is provided in Fig. 6. Obviously, after introducing M_e and M_d, the loss convergence curve in network training is smoother and the amplitude of oscillation is reduced, indicating that the training process is more stable.

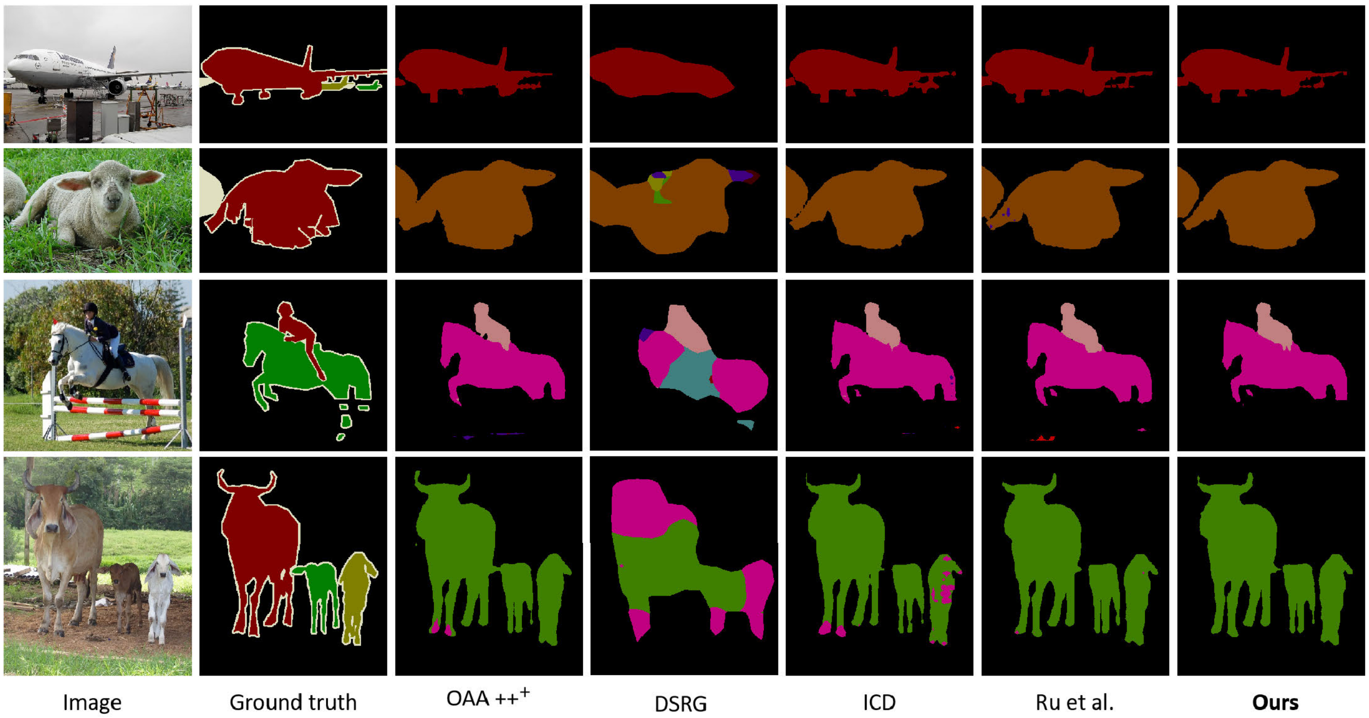


Fig. 5. The visualization of the semantic segmentation performance of our method. Some state-of-the-art weakly-supervised semantic segmentation methods in recent years were compared with our method, and experiments were implemented on PASCAL VOC 2012 dataset, with four typical categories selected to show the effect.

TABLE II

Quantitative Comparison of Weakly-Supervised Semantic Segmentation Methods on 21 Categories From PASCAL VOC 2012 Dataset

Methods	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mIoU
MIL+seg [49]	79.6	50.2	21.6	40.9	34.9	40.5	45.9	51.5	60.6	12.6	51.2	11.6	56.8	52.9	44.8	42.7	31.2	55.4	21.5	38.8	36.9	42
SEC [50]	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	75.2	22.1	53.5	28.3	65.8	57.8	62.3	52.5	32.5	62.6	32.1	45.4	45.3	50.7
AdvErasing [51]	83.4	71.1	30.5	72.9	41.6	55.9	63.1	60.2	74.0	18.0	66.5	32.4	71.7	56.3	64.8	52.4	37.4	69.1	31.4	58.9	43.9	55.0
AffinityNet [7]	88.2	68.2	30.6	81.1	49.6	61.0	77.8	66.1	75.1	29.0	66.0	40.2	80.4	62.0	70.4	73.7	42.5	70.0	42.6	**68.1**	51.6	61.7
SEAM [11]	88.8	68.5	33.3	**85.7**	40.4	67.3	78.9	**76.3**	81.9	29.1	75.5	**48.1**	79.9	73.8	71.4	**75.2**	48.9	79.8	40.9	58.2	53.0	64.5
Ours	**89.0**	**73.4**	**35.9**	78.9	**52.9**	**67.3**	**82.2**	72.4	**86.2**	**30.3**	**78.4**	40.2	**80.5**	**79.1**	**80.8**	66.0	**57.7**	**81.3**	**57.6**	62.0	**55.0**	**67.3**

TABLE III

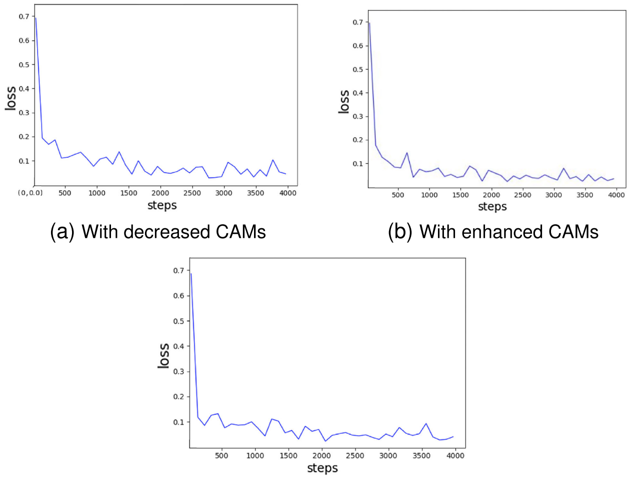
Ablation Study of the Attention Transfer Module Components

Introducing M_e	Introducing M_d	mIoU
✘	✘	49.12
✔	✘	49.80
✘	✔	49.54
✔	✔	50.86

TABLE IV

Ablation Study of the Focused-unFocused Module Components

+ Multi-scale atrous	+ context association	+ Attention	mIoU
✘	✘	✘	47.82
✔	✘	✘	48.50
✘	✔	✘	48.64
✔	✔	✔	49.12



(a) With decreased CAMs

(b) With enhanced CAMs

(c) With both adjusted CAMs

Fig. 6. The visualization of loss convergence curves in ablation study for attention transfer module. The horizontal axis denotes steps and the vertical axis denotes loss.

In addition, we also performed an ablation study to evaluate the effect of hyper-parameters λ_s and λ_{bs} on the quality of CAMs generation. The quality of generated CAMs as pseudo-labels is evaluated by comparing with semantic segmentation ground truth labels, as illustrated in Table V. Based on the comparison, we found that setting hyper-parameters λ_s and λ_{bs} to 0.5 each resulted in the highest quality of

TABLE V

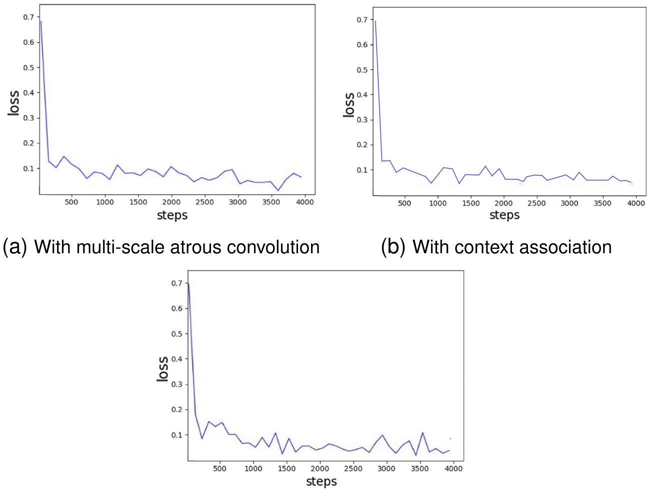
Ablation Study of the Hyper-Parameters λ_s and λ_{bs} for Pseudo-Labels Quality

λ_s	λ_{bs}	mIoU
0.1	0.1/0.2/0.5/1	50.79/50.79/50.81/50.78
0.2	0.1/0.2/0.5/1	50.79/50.82/50.83/50.80
0.5	0.1/0.2/0.5/1	50.81/50.84/50.86/50.83
1	0.1/0.2/0.5/1	50.79/50.81/50.84/50.85

TABLE VI

Ablation Study of the Different Attention Mechanisms for Pseudo-Labels Quality

+ Channel attention	+ Spatial attention	mIoU
✘	✘	47.82
✔	✘	48.69
✘	✔	48.12
✔	✔	48.90



(a) With multi-scale atrous convolution

(b) With context association

(c) Full model

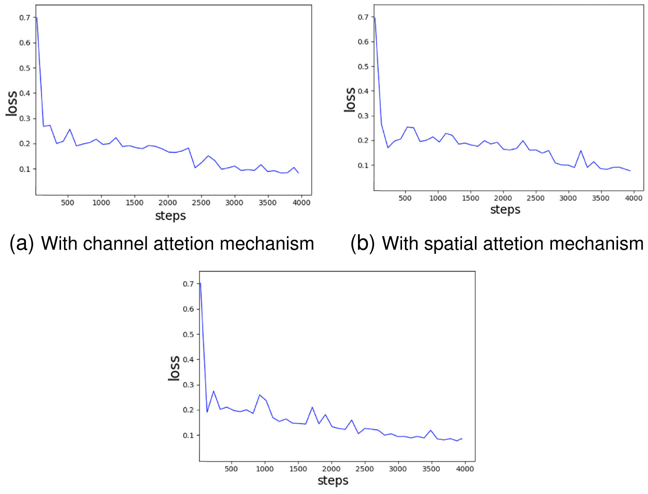
Fig. 7. The visualization of loss convergence curves in ablation study for multi-scale atrous convolution and context association submodule. The horizontal axis denotes steps and the vertical axis denotes loss.

generated CAMs. Therefore, we used this configuration in our training.

2) Effect of the Focused-unFocused Module: The focused-unfocused module contains three components, i.e., attention mechanisms, multi-scale atrous convolution and context association submodule, where attention mechanisms consist of channel attention and spatial attention. We first performed an ablation study on multi-scale atrous convolution and context association submodule to illustrate their contributions to the framework, as shown in Table IV, followed by a separate ablation study on the two attention mechanisms to illustrate the contribution of each adopted attention, as shown in Table VI.

As shown in Table IV, the multi-scale atrous convolution and context association submodule obviously improve the quality of generated pseudo-labels, with the loss convergence curves shown in Fig. 7. Note that, the base model of this ablation study only used M_o generated from original images.

The ablation study on the introducing of different attention mechanisms is illustrated in Table VI. From the results, we can see the channel attention and spatial attention both improve the quality of generated pseudo-labels, with the loss convergence curves shown in Fig. 8. Note that, the base model of this ablation study only used M_o generated from original images, without the introducing of multi-scale atrous convolution and context association submodule.



(a) With channel attetion mechanism

(b) With spatial attetion mechanism

(c) With both mechanisms

Fig. 8. The visualization of loss convergence curves in ablation study for attention mechanisms. The horizontal axis denotes steps and the vertical axis denotes loss.

V. Conclusion

To improve the performance of weakly-supervised segmentation, a novel method is proposed, which generates class activation/attention maps (CAMs) containing sufficient semantic information as pseudo-labels for the semantic segmentation training without pixel-level labels. In this method, the attention-transfer module is designed to guide the CAMs focus on both salient and inconspicuous regions of the targets, with sufficient semantic information extracted for pseudo-labels. A pixel relevance focused-unfocused module has also been developed for better integrating contextual information, with both attention mechanisms employed to extract focused relevant pixels and multi-scale atrous convolution employed to expand receptive field for establishing distant pixel connections. The proposed method has been experimentally demonstrated to achieve competitive performance in weakly-supervised segmentation on PASCAL VOC 2012 dataset, and even outperforms many saliency-joined methods. This method that employs neural networks to generate pseudo-labels, offers an alternative solution for obtaining annotations in the field of semantic segmentation, leading to substantial savings in human labor costs. It also suggests a potential way for future developments in the AIGC field, expanding AIGC technology to a broader range of tasks through the generation of auxiliary information.

References

[1] S.-J. Zhang, J.-H. Pan, J. Gao, and W.-S. Zheng, “Semi-supervised action quality assessment with self-supervised segment feature recovery,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6017–6028, Sep. 2022.

[2] Z. Li, H. Tang, Z. Peng, G.-J. Qi, and J. Tang, "Knowledge-guided semantic transfer network for few-shot image recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, 2023, doi: 10.1109/TNNLS.2023.3240195.

[3] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[4] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[5] L. Ru, H. Zheng, Y. Zhan, and B. Du, "Token contrast for weakly-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3093–3102.

[6] C. Song, W. Ouyang, and Z. Zhang, "Weakly supervised semantic segmentation via box-driven masking and filling rate shifting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15996–16012, Dec. 2023.

[7] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4981–4990.

[8] J. Ahn, S. Cho, and S. Kwak, "Weakly supervised learning of instance segmentation with inter-pixel relations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2209–2218.

[9] Y. Hao et al., "Attention in attention: Modeling context correlation for efficient video classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 7120–7132, Oct. 2022.

[10] Z. Chen, T. Wang, X. Wu, X.-S. Hua, H. Zhang, and Q. Sun, "Class re-activation maps for weakly-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 969–978.

[11] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12275–12284.

[12] Y. Sun and Z. Li, "SSA: Semantic structure aware inference for weakly pixel-wise dense predictions without cost," 2021, *arXiv:2111.03392*.

[13] Z. Li, Y. Sun, L. Zhang, and J. Tang, "CTNet: Context-based tandem network for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9904–9917, Dec. 2022.

[14] Y. Sun et al., "Singular value fine-tuning: Few-shot segmentation requires few-parameters fine-tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 37484–37496.

[15] S. Xu, D. Liu, and Z. Xiong, "E2I: Generative inpainting from edge to image," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1308–1322, Apr. 2020.

[16] F. Qian, Y. He, Y. Yue, Y. Zhou, B. Wu, and G. Hu, "Improved low-rank tensor approximation for seismic random plus footprint noise suppression," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–19, 2023, doi: 10.1109/TGRS.2023.3243831.

[17] F. Qian et al., "Unsupervised seismic footprint removal with physical prior augmented deep autoencoder," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–20, 2023, doi: 10.1109/TGRS.2023.3277973.

[18] S. Li, Z. Yu, M. Xiang, and D. Mandic, "Reciprocal GAN through characteristic functions (RCF-GAN)," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2246–2263, Feb. 2023.

[19] S. Li, Z. Yu, M. Xiang, and D. Mandic, "Reciprocal adversarial learning via characteristic functions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 217–228.

[20] Y. Liu, Y. Zhou, K. Yang, and X. Wang, "Unsupervised deep learning for IoT time series," *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14285-14306, 2023, doi: 10.1109/JIOT.2023.3243391.

[21] Y. Zhou, X. Song, Y. Zhang, F. Liu, C. Zhu, and L. Liu, "Feature encoding with autoencoders for weakly supervised anomaly detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 6, pp. 2454–2465, Jun. 2022.

[22] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[23] L. C. Chen, G. Papandreou, and I. Kokkinos, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Jun. 2017.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[25] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, "An empirical study of spatial attention mechanisms in deep networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6688–6697.

[26] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

[27] P. Wang et al., "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1451–1460.

[28] T. Y. Lin, P. Dollàr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[30] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7014–7023.

[31] Q. Hou, P. Jiang, Y. Wei, and M.-M. Cheng, "Self-erasing network for integral object attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/c042f4db68f23406c6cecf84a7ebb0fe-Paper.pdf

[32] W. Shimoda and K. Yanai, "Self-supervised difference detection for weakly-supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5208–5217.

[33] J. Fan, Z. Zhang, T. Tan, C. Song, and J. Xiao, "CIAN: Cross-image affinity net for weakly supervised semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 10762–10769.

[34] T. Zhang, G. Lin, W. Liu, J. Cai, and A. C. Kot, "Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Switzerland: Springer, 2020, pp. 663–679.

[35] G. Sun, W. Wang, J. Dai, and L. Van Gool, "Mining cross-image semantics for weakly supervised semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 347–365.

[36] L. Chen, W. Wu, C. Fu, X. Han, and Y. Zhang, "Weakly supervised semantic segmentation with boundary exploration," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 347–362.

[37] Y.-T. Chang, Q. Wang, W.-C. Hung, R. Piramuthu, Y.-H. Tsai, and M.-H. Yang, "Weakly-supervised semantic segmentation via sub-category exploration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8991–9000.

[38] X. Wang, S. Liu, H. Ma, and M.-H. Yang, "Weakly-supervised semantic segmentation by iterative affinity learning," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1736–1749, Jan. 2020.

[39] N. Araslanov and S. Roth, "Single-stage semantic segmentation from image labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4253–4262.

[40] J. Fan, Z. Zhang, C. Song, and T. Tan, "Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4283–4292.

[41] D. Zhang, H. Zhang, J. Tang, X.-S. Hua, and Q. Sun, "Causal intervention for weakly-supervised semantic segmentation," in *Proc. NIPS*, 2020, pp. 655–666.

[42] P.-T. Jiang, L.-H. Han, Q. Hou, M.-M. Cheng, and Y. Wei, "Online attention accumulation for weakly supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 7062–7077, Oct. 2022.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[44] Y. Su, R. Sun, G. Lin, and Q. Wu, "Context decoupling augmentation for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6984–6994.

[45] L. Ru, Y. Zhan, B. Yu, and B. Du, "Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16846–16855.

[46] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Sys. (NIPS)*, vol. 34, Dec. 2021, pp. 12077–12090.

[47] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 991–998.

[48] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[49] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1713–1721.

[50] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 695–711.

[51] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1568–1576.

Zhen Qin (Member, IEEE) received the Ph.D. degree from the University of Electronic Science and Technology of China (UESTC) in 2012. He is currently a Professor with the School of Information and Software Engineering, UESTC. He was a Visiting Scholar with the Department of Electrical Engineering and Computer Science, Northwestern University. His research interests include data fusion analysis, mobile social networks, wireless sensor networks, and image processing.

Yujie Chen received the bachelor's degree from the University of Electronic Science and Technology of China (UESTC) in 2020, where she is currently pursuing the master's degree with the School of Information and Software Engineering. Her research interests include semantic segmentation and instance segmentation.

Guosong Zhu received the B.S. degree (Hons.) in software engineering from the University of Electronic Science and Technology of China (UESTC) in 2019, where he is currently pursuing the Ph.D. degree with the School of Information and Software Engineering. He has served as a principal investigator in major national research instrument development projects. His research interests include image processing and multidimensional reconstruction.

Erqiang Zhou received the Ph.D. degree from the Dublin Institute of Technology (currently Technological University Dublin) in 2011. He is an Associate Professor with the School of Information and Software Engineering, University of Electronic Science and Technology of China (UESTC), where he joined in 2012. His research interests are mainly in the fields of password security, compiler optimization, and intelligent systems in education.

Yingjie Zhou (Member, IEEE) received the Ph.D. degree from the School of Communication and Information Engineering, University of Electronic Science and Technology of China (UESTC), China, in 2013. He is currently an Associate Professor with the College of Computer Science, Sichuan University (SCU), China. He was a Visiting Scholar with the Department of Electrical Engineering, Columbia University, New York. His current research interests include behavioral data analysis, machine learning, network management, and resource allocation. He received the Best Paper Awards at IEEE HPCC and IEEE MMSP in 2022. He has served as the Program Vice-Chair for IEEE HPCC; the Local Arrangement Chair for IEEE BMSB; and a TPC Member for many major IEEE conferences, such as GLOBLECOM, ICC, ITSC, MSN, and VTC.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree in electrical engineering from Hunan University, Changsha, China, and the M.S. and Ph.D. degrees in electrical engineering from Tufts University, Medford, MA, USA.

He is a Professor with the Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, machine learning, and multimedia security.

Dr. Zhou is a fellow of the Society of Photo-Optical Instrumentation Engineers (SPIE). He was recognized as one of the highly cited researchers in 2020, 2021, and 2023. He serves as an Associate Editor for IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Ce Zhu (Fellow, IEEE) received the B.S. degree in electronic and information engineering from Sichuan University, Chengdu, China, in 1989, and the M.Eng. and Ph.D. degrees in electronic and information engineering from Southeast University, Nanjing, China, in 1992 and 1994, respectively.

He was a Post-Doctoral Researcher with The Chinese University of Hong Kong, Hong Kong, in 1995; the City University of Hong Kong, Hong Kong; and The University of Melbourne, Melbourne, VIC, Australia, from 1996 to 1998. He was with Nanyang Technological University, Singapore, from 1998 to 2012, for 14 years, where he was a Research Fellow, a Program Manager, an Assistant Professor, and then promoted to an Associate Professor in 2005. Since 2012, he has been with the University of Electronic Science and Technology of China, Chengdu, as a Professor. His research interests include video coding and communications, video analysis and processing, 3D video, and visual perception and applications.

Dr. Zhu was a co-recipient of multiple paper awards at international conferences, including most recently the Best Demo Award in IEEE MMSP 2022 and the Best Paper Runner-Up Award in IEEE ICME 2020. He has served on the editorial boards for a few journals, including an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON BROADCASTING, and IEEE SIGNAL PROCESSING LETTERS; an Editor for IEEE COMMUNICATIONS SURVEYS AND TUTORIALS; and an Area Editor for *Signal Processing: Image Communication*. He has also served as a Guest Editor for a few special issues in international journals, including a Guest Editor for IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. He was an APSIPA Distinguished Lecturer, from 2021 to 2022, and an IEEE Distinguished Lecturer of Circuits and Systems Society, from 2019 to 2020.