Chinese Journal of Electronics
Vol.32, No.4, July 2023

Attention Guided Enhancement Network for Weakly Supervised Semantic Segmentation

ZHANG Zhe[1,2], WANG Bilin[1,2], YU Zhezhou[1,2], and ZHAO Fengzhi[1]

(1. *College of Computer Science and Technology, Jilin University, Changchun 130012, China*)
(2. *Key Laboratory for Symbol Computation and Knowledge Engineering of National Education Ministry, Changchun 130012, China*)

Abstract — Weakly supervised semantic segmentation using only image-level labels is critical since it alleviates the need for expensive pixel-level labels. Most cutting-edge methods adopt two-step solutions that learn to produce pseudo-ground-truth using only image-level labels and then train off-the-shelf fully supervised semantic segmentation network with these pseudo labels. Although these methods have made significant progress, they also increase the complexity of the model and training. In this paper, we propose a one-step approach for weakly supervised image semantic segmentation—attention guided enhancement network (AGEN), which produces pseudo-pixel-level labels under the supervision of image-level labels and trains the network to generate segmentation masks in an end-to-end manner. Particularly, we employ class activation maps (CAM) produced by different layers of the classification branch to guide the segmentation branch to learn spatial and semantic information. However, the CAM produced by the lower layer can capture the complete object region but with many noises. Thus, the self-attention module is proposed to enhance object regions adaptively and suppress irrelevant object regions, further boosting the segmentation performance. Experiments on the Pascal VOC 2012 dataset demonstrate that AGEN outperforms alternative state-of-the-art weakly supervised semantic segmentation methods exclusively relying on image-level labels.

Key words — Weakly-supervised learning, Semantic segmentation, Convolutional neural networks, Self-attention mechanism, End-to-end.

I. Introduction

Semantic segmentation [1], [2] aims at predicting the classification results of every pixel in an image and is one of the most critical tasks in the field of computer vision. Deep convolutional neural networks (DCNNs) have made significant improvements in the performance of fully supervised semantic segmentation in recent years. However, the performance of these fully supervised methods is heavily dependent on the availability of large-scale datasets with pixel-level labels for training, which results in large consumption of time and labor. Many works [3]–[8] are devoted to supervised semantic segmentation with weaker annotations obtained in a more accessible and cost-effective manner to alleviate the burden of expensive data gathering using pixel-level labels.

Among those mentioned above weakly supervised annotations, we devote attention to the most challenging semantic segmentation task, which adopts image-level annotations that can be collected in a great quantity most cheaply. Under the supervision of pixel-level annotation, the semantic segmentation method can learn the reliable boundary information of objects and their relationship in the image. By contrast, the image-level annotations demonstrate the presence of specific-class objects in the map without providing information on their border and location, which are essential to training semantic segmentation methods. As a result, the first critical task is to create a correspondence between the image-level label and the pixel location information of the image. To this end, a popular approach, class activation maps (CAM) [9], was proposed to localize the most discriminative regions from a pre-trained image classification network. However, only the sparse and incomplete areas with the best recognition ability may be obtained since these class activation maps come from the classification network and only fo-

Manuscript Received July 10, 2021; Accepted Oct. 21, 2021. This work was supported by the Development Project of Jilin Province of China (20200801033GH, 2020122328JC), the Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University (20180520017JH), and the Fundamental Research Funds for the Central Universities, JLU.
 DOI:10.23919/cje.2021.00.230

cus on classification accuracy rather than object integrity. Thus there is a significant disparity between the pseudo-ground truth produced by classification networks and the pixel-level labels. To narrow this gap, most current weakly supervised methods of image semantic segmentation adopt a two-step framework, which firstly adopts class activation maps as initial seeds to generate high-quality pseudo-pixel-level labels, and then fed these pseudo labels as ground-truth into off-the-shelf fully supervised semantic segmentation network [2] for training. Although these methods have made significant progress, they result in an increase in model and training complexity, such as additional training cycles [3], [4], numerous models training [5], [6], and off-the-shelf saliency methods [7], [8]. While early methods [10], [11] often adopt one step framework consisting of one network structure. Most of these methods perform semantic segmentation within the multi-instance learning (MIL) framework and combine with other restricted optimization methodologies, which is easy to implement without additional bells and whistles. The segmentation accuracy of these approaches, on the other hand, is significantly below that of the fully supervised method.

This paper presents an efficient one-step method for image-level weakly supervised semantic segmentation—attention-guided enhancement network (AGEN) that can be trained end-to-end. The proposed AGEN consists mostly of two parallel branches: classification and segmentation. The classification branch is utilized to produce pseudo-pixel-level labels with the supervision of image-level labels. The segmentation branch is adopted to produce semantic segmentation results under the supervision of pseudo-pixel-level labels. Lower layers of the neural network can extract more spatial information (e.g., edge and boundary), and upper layers have more abstract semantic information [2], [12], [13]. We propose attention guided module (AGM), which can use class activation maps produced in lower layers to guide segmentation branch learn spatial information and produced by upper layers to guide segmentation branch learn semantic information in a bottom-up manner. Although the class activation maps produced lower layers have more spatial information, there are also generate a large scale of noises which harming the performance of segmentation. To address this issue, we present the context attention module (CoAM), which captures context-reliance between class-aware feature maps produced by the different layers of the network model, which enable adaptively enhance object regions and suppress irrelevant object regions for further improve the performance of segmentation.

Finally, we optimize the proposed Attention guided enhancement network by minimizing the joint loss function and produce segmentation results in an end-to-end fashion. Additionally, we demonstrate in the experiments that our approach is validated against the challenging PASCAL VOC 2012 segmentation benchmark [14] for image-level weakly supervised image semantic segmentation.

The following are the major contributions of our work:

- We present a one-step strategy for weakly semantic segmentation using an attention-guided enhancement network (AGEN) that can be trained end-to-end.
- We introduce attention guided module (AGM) and context attention module (CoAM) in the AGEN. AGM can guide the network to learn spatial and semantic information in a bottom-up fashion, and CoAM can optimize AGM to improve segmentation performance further.
- On the PASCAL VOC 2012 benchmark, our method relies entirely on image-level labels, yielding state-of-the-art results for weakly supervised semantic segmentation and outperforming other weakly supervised algorithms that use two-step structures or extra information.

II. Related Works

Deep convolutional neural networks [15] (DCNN) have achieved excellent results in a variety of computing tasks [1], [12], [13], [16], but training them requires a considerable number of fully annotated datasets. Significantly, the task of semantic segmentation [1], [2] demands fully supervised pixel-level labels, which are costly to obtain. In order to reduce the burden of obtaining costly pixel-level labels, a range of methods for semantic segmentation using various types of weakly supervised labels have been proposed [17]–[23]. This paper discussed weakly-supervised semantic segmentation using just image-level labels that are the most cost-effective to collect.

Recently, most image-level weakly supervised semantic segmentation algorithms based on CAM [9] can generate initial object seed areas under the supervision of image-level labels and use the regions to generate pseudo-ground truth for training the semantic segmentation network. Kolesnikov *et al.* [20] proposed to integrate three principles, including the seed, expand, and constrains (SEC), into a unified framework to train semantic segmentation models. However, the SEC concentrates exclusively on small and sparse seed regions of objects, which provide insufficient supervision information for training the semantic segmentation network. A number of methods have been proposed in recent years to tackle this problem, which may be classified into

static and dynamic masks respectively.

The main feature of this method is that the pseudo label will not change with the training of the segmentation network. A series of erasing-based approaches [21]–[23] are proposed to expand the object seed regions, in which Wei *et al.* [3] is the first to propose an adversarial erasing strategy that iteratively erases the most discriminative regions and then drove the classification network to find new object regions, Chaudhry *et al.* [21] utilizes the saliency detector [24] to find new salient object regions in erasing manner, Li *et al.* [22] proposes guided attention inference networks (GAIN) to provide self-guidance on the classification for improving the erasing strategy, and Hou *et al.* [23] proposes self-erasing network (SeeNet) that contained two self-erasing strategies to suppress object regions spread to background regions [1], [5], [8], [9], [25]. Jiang *et al.* [26] offer the online attention accumulation (OAA) method for gradually promoting the generation of whole object regions by accumulating attention maps with different classification network training epochs SEAM proposed by Wang *et al.* [27] to improve the prediction consistency for given the image with various transformed, boost the classification network to produce high-quality object seed regions.

For the dynamic mask, generating pseudo-pixel-level labels during training a fully supervised semantic segmentation network. That is while optimizing the semantic segmentation network, and the pseudo-ground-truth are constantly updated [4]. Deep seed region growing (DSRG) [28] is a method that combines traditional algorithms with deep convolutional neural networks to dynamically and gradually extend the object seed regions during the training of its semantic segmentation network. To enhance segmentation performance, Shimoda *et al.* [29] propose self-supervised difference detection (SSDD) to estimate the noises in the segmentation results produced by conditional random fields (CRF). Fan *et al.* [30] present a cross-image affinity module for capturing relationships between two independent images containing objects of the same class, hence giving extra information for incomplete object regions.

All the methods mentioned above require sophisticated multi-step training procedures for producing high-quality segmentation results. In contrast, we propose a one-step approach for image-level weakly semantic segmentation, which simplifies the process of generating segmentation results.

III. Approach

We will go through each component of the proposed attention guided enhancement network in detail in this section. Firstly, we give an overview of AGEN's end-to-end structure. Second, we introduce the AGM to boost segmentation branch performance. Then, the context attention module is presented to enhance the quality of class attention maps created by the AGM, hence enhancing the segmentation method performance even more. Finally, we give the details of AGEN loss.

1. Overview of AGEN structure

This paper proposes an end-to-end image-level weakly supervised semantic segmentation method, which enables dynamically generating class activation maps as supervision for a bottom-up guidance training segmentation network. As shown in Fig.1, the proposed attention-guided enhancement network mainly contains two parallel branches: classification branch A and segmentation branch B, respectively. Branch A shares the first two layers parameters of the backbone network with branch B and minimizes the joint loss function L ($L = L_{\text{class}} + L_{\text{seg}}$) to update all network parameters simultaneously during training, where L_{class} represents classification loss of branch A, and L_{seg} represents segmentation loss of branch B.

The classification branch A is used to produce the class activation map, which serves as pseudo-pixel-level labels, for the training segmentation branch. Similar to the previous methods [5], [21], we also adopt the class-aware feature maps produced by the last convolutional layer of the classification branch to generate object region maps, which proved by [31] identical to the production process of class activation maps proposed by [9]. Specifically, we adopt modified VGG-16 [32] as the backbone of our branch A, in which two 3×3 filters with 1024 channels and one 1×1 filter with C channels (where C is the number of foreground classes) replace the fully connected layers. Then, a global average pooling (GAP) operation is performed on the class-specific feature mappings to generate a tensor representing the map. Finally, the result of classification prediction is obtained by a sigmoid function, which is defined as follows:

$$p_c = \frac{1}{1 + \mathrm{e}^{\mathrm{GAP}(F_c)}} \tag{1}$$

where $c \in C$ represents the target category, F_c denoted the cth feature map from the last class-aware 1×1 convolutional layer.

For given an image I, we first input F_c into a ReLU layer, and then the class activation maps of target category c can be obtained as follows:

$$M_c = \mathrm{US}(\mathrm{ReLU}(F_c)) \tag{2}$$

where $\mathrm{US}(\cdot)$ denotes a feature map that has been upsampled to the same size as the input image I through bi-linear interpolation.

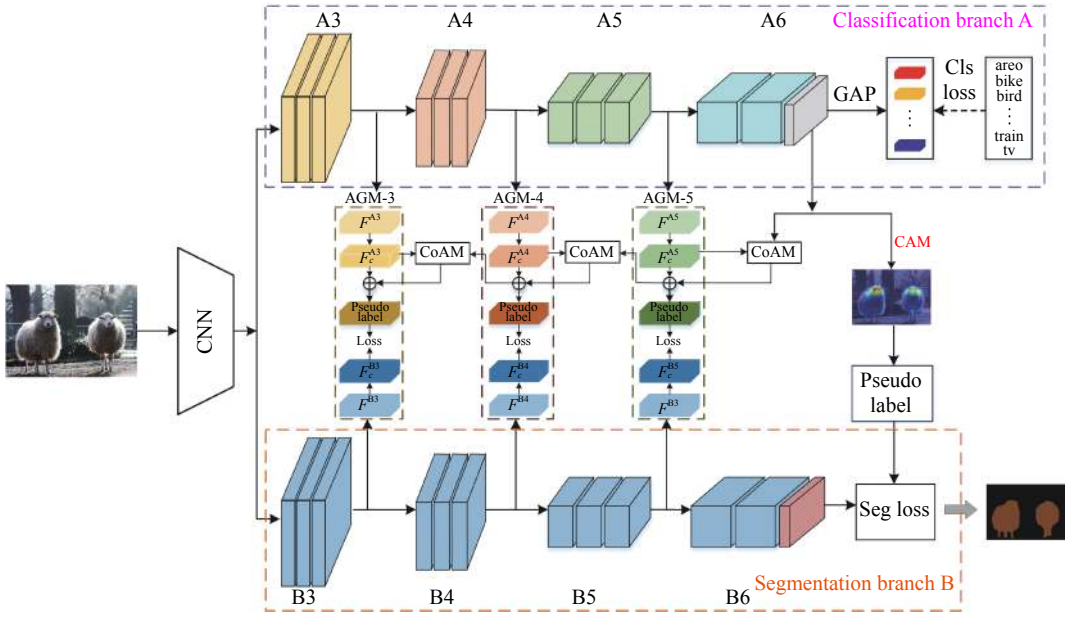


Fig. 1. The overall framework of the proposed attention guided enhancement network.

To make sure the prediction value of each pixel in I range from 0 to 1, the CAM produced by (2) are normalized as follows:

$$M_c^{\mathrm{f}} = \frac{M_c}{\max(M_c)} \tag{3}$$

Although M_c^{f} is optimized by minimizing the foreground target classification loss, many background pixels are crucial for the training segmentation branch. For obtaining a class activation map of background, we adopt a similar way in [5], which is defined as follows:

$$M^{\mathrm{b}}(i) = \{1 - \max_{c \in C} M_c^{\mathrm{f}}(i)\}^{\partial} \tag{4}$$

where $M_c^{\mathrm{f}}(i)$ denoted the prediction value of the foreground object class activation maps M^{f} for category c at position i, $\partial \geq 1$ is the hyper-parameter that can adjust the value of the pixel i background label. Then, the overall class activation maps $M^{\mathrm{f\text{-}b}}$ is obtained by concatenating M^{f} and M^{b}, and we retrieve the highly confident regions H by setting values of pixels more enormous than a threshold δ. Finally, the pixel-level pseudo labels of each training image are obtained for training the segmentation branch.

The segmentation branch B adopts the pseudo ground truth produced by classification branch A as supervision for training, which shares the first two layers of the classification branch, and the remain layers adopt the ResNet 38 [33] network architecture. Finally, optimize the whole network of our proposed approach by minimizing the classification loss and segmentation loss simultaneously, and then generate the segmentation result under the guidance of the attention guided module.

2. Attention guided module

Only adopting pseudo-pixel-level labels produced by the classification branch as supervision for the training segmentation branch cannot get the expected result. As we know, a classification network is usually to find a common semantic pattern for a specific class to identify the object in the image, so it produces the class activation maps only highlighting the most discriminative object regions, which are small and sparse. Lower layers of the neural network can extract more spatial information of the object (e.g., edge and boundary), and upper layers can acquire more abstract semantic information [2], [12], [13]. We propose AGM, as shown in Fig.1 (AGM-3, AGM-4 and AGM-5), which adopts a bottom-up mechanism. Using class activation maps produced by lower layers of classification branch to guide segmentation branch spatial information learning and using the upper layers to produce class activation maps as supervision for enhancing segmentation branch semantic information learning. The detail of the proposed AGM is described as follows:

AGM utilizes the intermediate feature maps produced by classification branch A to generate class activation maps as a guide for improving segmentation branch B. As shown in Fig.1, the intermediate feature maps ($F^{\mathrm{A}_3}, F^{\mathrm{A}_4}, F^{\mathrm{A}_5}$ and $F^{\mathrm{B}_3}, F^{\mathrm{B}_4}, F^{\mathrm{B}_5}$) produced by classification branch A and segmentation branch B, respectively, are fed into AGM-3, AGM-4, AGM-5. Particularly, each AGM also contains two branches: Classification branch C and segmentation branch D, as shown in Fig.2.

The classification branch C comprises cascaded layers: two 3×3 convolutional layers with 256 channels

and a 1 × 1 convolutional layer with C channels corresponding to the number of classes, followed by a global average pooling (GAP) and a sigmoid layer. The first 3 × 3 convolutional layer is used to unify the different numbers of the channel of the input feature maps (F^{A_3}, F^{A_4} and F^{A_5}), and AGM-3, AGM-4, AGM-5 share the last two convolutional layers of branch C. F_c^A is obtained by the last class-aware 1 × 1 convolutional layer of branch C. The segmentation branch D consists of three convolutional layers: two 3 × 3 convolutional layers with 512 channels and one 1 × 1 convolutional layer with C+1 channels (where C is the number of object categories and one background class). The first convolutional layer of branch D also adapts the different numbers of the channel in features maps F^B produced by branch B at different layers and the two last layers shared by AGM at different stages. The class-aware F_c^B produced by branch D, which is resized to the same size as input image I, generates segmentation maps S. Then, we use (2) and (3) with F_c^A as input to produce pseudo-pixel-level labels as supervision for training segmentation branch D. Finally, the AGM is optimized by minimizing the classification and segmentation loss of branches C and D, encouraging the attention guided enhancement network to learn spatial object information of the image in a bottom-up fashion.

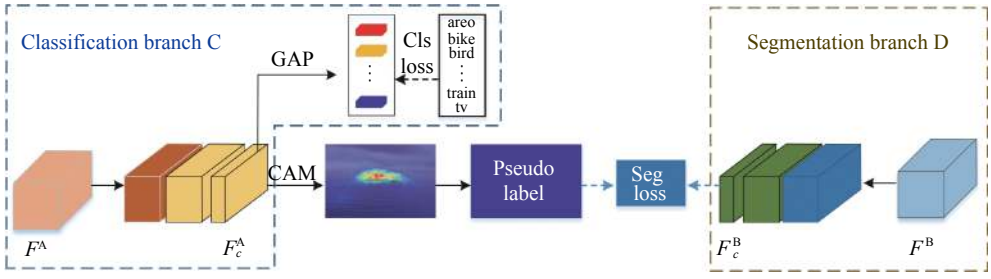


Fig. 2. Illustration of the proposed attention guided module.

3. Context attention module

Although the class activation maps produced by the lower layer have more spatial information of objects which can act as supervision to guide the segmentation branch to segment the complete object in the input image, some true negative regions are falsely highlighted, which harms the performance of the segmentation branch. To address this issue, we propose a context attention module (CoAM) that leverages the self-attention mechanism [34] to find dependencies between class-aware feature maps generated by the classification branch at various phases. Class-aware feature maps produced by the lower layer can capture the complete object region but with some noises, and produced by the upper layer only identify small and sparse discriminative object regions but which is reliable. As shown in Fig.1, we adopt CoAM in class-aware feature maps produced by different layers of classification branch in a top-down fashion, which is denoted as follows:

$$\boldsymbol{F}_c^{A_{n-1}} = \mathrm{CoAM}(\boldsymbol{F}_c^{A_{n-1}}, \boldsymbol{F}_c^{A_n}) \tag{5}$$

where $F_c^{A_{n-1}}$ and $F_c^{A_n}$ denoted class activation maps produced by adjacent layers of classification branch, in which $F_c^{A_n}$ is up-sampled to be the same size as $F_c^{A_{n-1}}$ through bi-linear interpolation.

The detail of CoAM operation are described as follows: As illustrated in Fig.3, the module takes the $F_c^{A_{n-1}} \in \mathbb{R}^{C\times H\times W}$ and $F_c^{A_n} \in \mathbb{R}^{C\times H\times W}$ as input. Then, two class-aware feature maps are further reshaped into $\mathbb{R}^{C\times N}$, where $N = H \times W$ is the number of the pixel location. Then, the context attention map $A \in \mathbb{R}^{N\times N}$ is obtained by performing the matrix product between the reshaped class-aware feature maps of $F_c^{A_{n-1}}$ and $F_c^{A_n}$, where each position score of context attention maps can be defined by

$$a_{i,j} = \exp((F_{c,i}^{A_{n-1}})^{\mathrm{T}} F_{c,j}^{A_n}) \tag{6}$$

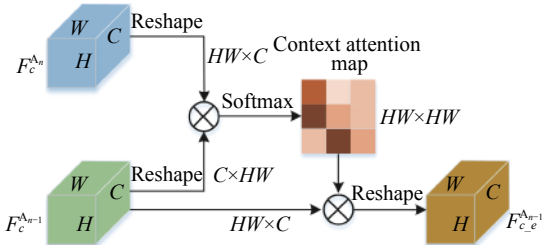


Fig. 3. Illustration of the proposed context attention module.

where $\{i, j\} \in \{1, 2, \ldots, N\}$ denoted the index of pixels position from reshaped class-aware feature maps $F_c^{A_{n-1}}$ and $F_c^{A_n}$. $a_{i,j}$ measures the affinity of reshaped $F_c^{A_{n-1}}$ and $F_c^{A_n}$ at positions i and j, in which the more similar feature of the two positions is enhanced, and the irrelated feature of them is suppressed. Then we normalize the $a_{i,j}$ by adopting softmax operation to ensure that the total of all the weights associated with a pixel is one unit:

$$\bar{a}_{i,j} = \frac{a_{i,j}}{\sum_{j=1}^{N} a_{i,j}} \tag{7}$$

To improve the original class-aware feature maps $F_c^{A_{n-1}}$, we reshape $F_c^{A_{n-1}}$ to $\mathbb{R}^{C\times N}$ and perform a matrix multiplication between $F_c^{A_{n-1}}$ and the transpose of A. Then, the result is reshaped back to $\mathbb{R}^{C\times H\times W}$ to get the enhanced $F_c^{A_{n-1}}$.

$$F_{c\_e,i}^{A_{n-1}} = \lambda \sum_{j=1}^{N} (\bar{a}_{i,j} \cdot F_{c,j}^{A_{n-1}}) \tag{8}$$

where λ is the weight parameter initialized to 0 and gradual-learned during the training process. Finally, as seen in Fig.1, we execute an element-by-element summing of the original class-aware feature mappings $F_c^{A_{n-1}}$ which can maintain the initial behavior of $F_c^{A_{n-1}}$:

$$F_{c,i}^{A_{n-1}} = F_{c\_e,i}^{A_{n-1}} + F_{c,i}^{A_{n-1}} \tag{9}$$

It can be informed that from (9), each position of final $F_c^{A_{n-1}}$ is the weighted sum of $F_{c\_e}^{A_{n-1}}$ and original class-aware feature maps $F_c^{A_{n-1}}$ at every position. Thus, CoAM can learn contextual information from a different stage of class-aware feature maps produced by the classification branch at each location, which enable adaptively enhance object regions and suppress irrelevant object regions.

4. Loss design of AGEN

The proposed attention guided enhancement network mainly contains two types of loss: classification loss and segmentation loss. For classification loss, we adopt Sigmoid cross-entropy loss as a multi-label classification loss function:

$$l_{\text{cls}}(p,q) = -\sum_{c=1}^{C} [q_c \log p_c + (1-q_c)\log(1-p_c)] \tag{10}$$

where q denoted the image-level classification labels, and p is the classification score of objects produced by (1). The proposed network got four classification scores (p^{A}, $p^{\text{AGM-5}}$, $p^{\text{AGM-4}}$, $p^{\text{AGM-3}}$), as shown in Fig.1, produced by classification branch A, AGM-5, AGM-4 and AGM-3 respectively. Then the classification loss is calculated as

$$\begin{aligned} L_{\text{cls}} = {} & l_{\text{cls}}\left(p^{A}, q\right) + l_{\text{cls}}\left(p^{\text{AGM}-5}, q\right) \\ & + l_{\text{cls}}\left(p^{\text{AGM}-4}, q\right) + l_{\text{cls}}\left(p^{\text{AGM}-3}, q\right) \end{aligned} \tag{11}$$

We employ balanced seeding loss [25], [28] for semantic segmentation, which considers the imbalanced distribution of confident foreground and background regions. Let C and $\hat{C}$ are denoted the set of foreground classes and the background class which is formulated as:

$$\begin{aligned} l_{\text{seg}}(H_c, S) = {} & -\frac{1}{\sum_{c\in C}|H_c|}\sum_{c\in C}\sum_{u\in H_c}\log S_{u,c} \\ & -\frac{1}{\sum_{c\in \hat{C}}|H_c|}\sum_{c\in \hat{C}}\sum_{u\in H_c}\log S_{u,c} \end{aligned} \tag{12}$$

where H_c represented a collection of object regions corresponding to class c and were generated by the classification branch. S represents the segmentation map produced by the segmentation branch, in which $S_{u,c}$ denoted the conditional probability of any label $c\in C$ at any location u of the segmentation map. $|\cdot|$ is the cardinality of pixels. The proposed method got four pairs H_c and S, which were produced by AGM-3, AGM-4, AGM-5 and overall network (classification branch A and segmentation branch B), as shown in Fig.1. Thus, the total segmentation loss is defined by

$$\begin{aligned} L_{\text{seg}} = {} & l_{\text{seg}}\left(H_c{}^{\text{AGM}-3}, S^{\text{AGM}-3}\right) + l_{\text{seg}}\left(H_c{}^{\text{AGM}-4}, S^{\text{AGM}-4}\right) \\ & + l_{\text{seg}}\left(\text{H}_\text{c}{}^{\text{AGM}-5}, \text{S}^{\text{AGM}-5}\right) + l_{\text{seg}}\left(\text{H}_\text{c}{}^{\text{Cla}}, \text{S}^{\text{SeB}}\right) \end{aligned} \tag{13}$$

We also use L_{boundary} boundary constrain loss [20], [25], [28] to encourage the segmentation maps matching with the object boundary which is the mean KL-divergence between outputs of the segmentation network and the conditional random field [35] defined as

$$L_{\text{boundary}} = \frac{1}{N}\sum_{u=1}^{N}\sum_{c=1}^{C+1} \boldsymbol{R}_{u,c}(I, S^{\text{SeB}})) \log \frac{\boldsymbol{R}_{u,c}(I, S^{\text{SeB}})}{S_{u,c}^{\text{SeB}}} \tag{14}$$

where I is the input image and $R_{u,\,c}\,(I,\,S^{\text{SeB}})$ denoted the output of fully connected CRF, in which S^{SeB} represent the segmentation mask produced by segmentation branch. Finally, the whole parameters of the network are updated at the same time by minimizing the total loss function L_{total}:

$$L_{\text{total}} = L_{\text{cls}} + L_{\text{seg}} + L_{\text{boundary}} \tag{15}$$

IV. Experiments

1. Experimental setup

1) Dataset and evaluation metrics

Our method was tested on the difficult PASCAL VOC 2012 image segmentation benchmark [14], which included 21 semantic class labels, 20 foreground categories and one background category. With 1464, 1449, and 1456 images, the dataset is divided into three parts: training (train), validation (val), and testing (test). Following the same strategy as earlier weakly supervised semantic segmentation approaches [3], [25], [28], we used an extended training set of 10582 images supplied

by [36] to train our network. In this experiment, we use just image-level labels as supervision for training the network. We use the mean intersection over union to evaluate the proposed approach's performance across all image classes, including the background class. For an in-depth analysis of the method experiment, only the validation set for which the ground truth is available is used. We compare our proposed method to existing state-of-the-art approaches in both the validation and test sets.

2) Implementation details

The proposed attention guided enhancement network is implemented on the publicly available PyTorch [37] framework. The training images are randomly resized within the range of [321, 481] and then cropped to 321 × 321 as input images of the network. We use the pre-trained weights on Image-Net [38] and randomly initialize the remaining parameters of the proposed method. The entire parameters of AGEN are then fine-tuned on the challenging PASCAL VOC 2012 images dataset [14] with the initial learning rate of 0.001 (0.01 for the attention guided module), and the learning rate is decreased by a factor of 10 after every ten epochs. The network is trained using a stochastic gradient descent (SGD) optimizer with mini-batch, and the training is terminated after 35 epochs. The batch size is set to 15, and the weight decay parameter is set to 5E−4. The parameter of SGD optimizer momentum is set to 0.9. To obtain the reliable object regions $M^{\text{f-b}}$ based on class-aware feature maps produced by classification branches, we set the parameters ∂ in (4) to 8. The threshold δ is set to 0.4 for AGM-3 and AGM-4, choosing the pixels belonging to the top 40% of the largest value for each class, and the threshold δ of AGM-5 and classification branch A is set to 0.3. Finally, during training, any unassigned and conflicting pixels are disregarded. During training, all the classification branches and segmentation branches update the entire attention guided enhancement network. During testing, we utilize the segmentation branch B to generate semantic segmentation results. All the experiments are implemented on a single NVIDIA GeForce RTX 2080Ti GPU with 11 GB memory.

2. Ablation studies

1) Effectiveness of proposed AGEN

We implement a series of ablation experiments using the Pascal VOC 2012 val dataset in various settings to assess the effectiveness of each element of the attention guided enhancement network. In our experiments, the "baseline" denoted the proposed end-to-end network without attention guided module and context attention module, just containing the classification branch A and segmentation branch B. Table 1 shows the performance of mIoU scores, which "OA" denoted add only one AGM on the baseline (i.e., AGM-4) can obtain a performance gain of nearly 3%. The main improvement is that the OA module can extract more spatial information of the object from the lower layer of the network and generate pseudo ground truth to guide segmentation learning spatial information. The "SC" means adding a CoAM based on the OA, which can capture the context-dependence between class-aware feature maps generated by classification branch A and AGM-4 respectively, so as to enhance the pixels of the object region and suppress the noise of the class-aware feature maps and get more than 1.5% performance improvement compared to the OA. The "MA" denotes adding multi-AGM on the baseline (i.e., AGM-3, AGM-4, AGM-5), and "AC" denotes CoAM is adopted based on MA in accumulated fashion, which improves the performance up to 58.5% and 63.1% mIoU on the PASCAL VOC 2012 val set respectively. The performance of MA and AC achieve 2.9% and 5.6% improvement respectively compared with the OA and SC adopted at the single stage of the network, which evaluates the effectiveness of adopting AGM and CoAM at the multi-stage of the network.

Table 1. The ablation experiments for each part of AGEN

Baseline	OA	SC	MA	AC	mIoU(%)
√					52.8
√	√				55.6
√	√	√			57.3
√	√		√		58.5
√	√		√	√	63.1

2) Analysis of AGM at different stages

To evaluate the effectiveness of the AGM at different stages in the network, we additionally implement a series of ablation experiments. As shown in Table 2, we only adopt AGM in the proposed network at the lowest stage (AGM-3) and at the upper stage (AGM-5), which achieved 0.5% and 1.1% performance improvement compared with baseline, respectively. It can be observed that the performance gain of the proposed network by add AGM-3 or AGM-5 has not been significantly improved compared with only adopted AGM-4 (2.8%) in the network. Because class activation maps

produced by the lower layer have more spatial information, it also brings much noise harming the performance of segmentation. Moreover, CAM produced by the upper layer extracts more semantic information but is insufficient to guide the network to generate high-quality segmentation results, as shown in Fig.4. We also adopt different multi-AGM in our network, the performance of the network gets improved compared to single-AGM, as shown in Table 2. Among these groups, we note that including AGM at three stages into the proposed network produces the most outstanding results, increasing performance by up to 58.5 percent mIoU.

Table 2. The ablation experiments for AGM at different stages

Baseline	AGM-3	AGM-4	AGM-5	mIoU(%)
√				52.8
√	√			53.3
√		√		55.6
√			√	53.9
√	√	√		56.3
√	√		√	55.7
√		√	√	56.6
√	√	√	√	58.5

3) Comparison of CoAM at different stages

We conducted experiments to evaluate adding accumulated CoAM base on the MA at different stages on the segmentation performance. Table 3 shows that adopts CoAM in the upper layer (i.e., AGM-5) of the network gives a performance gain of 0.3%. While incorporating accumulated CoAM into AGM-4 and AGM-3 results in a considerable performance gain of 2.2% and 4.6% mIoU, respectively, compared to MA. As shown in Fig.4, the CoAM may boost significant object regions while suppressing irrelevant ones, demonstrating the efficiency of the CoAM and emphasizing the need to merge AGM and CoAM to direct the network toward producing high-quality segmentation results.

3. Comparsion with state-of-the-arts

This section compares our approach to various state-of-the-art algorithms for image-level weakly supervised semantic segmentation in terms of mIoUs using the PASCAL VOC 2012 validation and test datasets. Firstly, we detailedly compare the proposed method with other previous end-to-end state-of-the-arts, and the results are presented in Table 4. Although there are many various weakly supervised semantic segmentation methods, to the best of our knowledge, only just three approaches of this task adopt an end-to-end architecture with only image-level supervision, which is EM-Adapt [10], CRF-RNN [11], and RRM [39] respectively. It

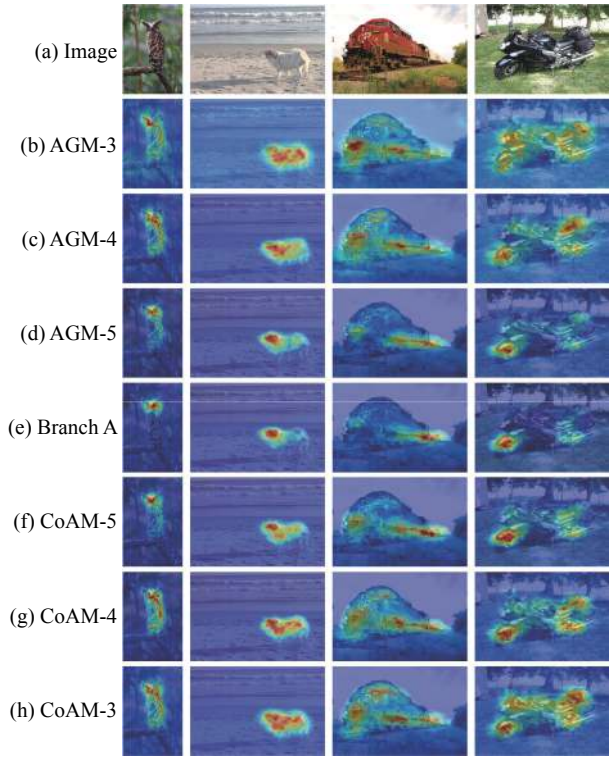


Fig. 4. Visual comparison of class activation maps produced by the proposed method at different stages.

Table 3. The ablation experiments for CoAM at different stages

MA	CoAM-5	CoAM-4	CoAM-3	mIoU(%)
√				58.5
√	√			58.8
√	√	√		60.7
√	√	√	√	63.1

can be observed from Tabel 4 that our AGEN outperforms the early end-to-end method (EM-Adapt and CRF-RNN) by large margins in terms of mIoUs on every class of the PASCAL VOC dataset. EM-Adapt utilizes the Expectation-Maximization algorithm to optimize the network, and CRF-RNN fuses three different computation processes into a segmentation network by a CRF as the recurrent network. The latest end-to-end method (RRM) gets the best performing one which is based on small and spares object regions and adds a shallow loss function to update the network. However, our end-to-end AGEN enhances performance by over 0.5% and 0.4% mIoUs on val and test datasets compared with the best performance of the approach. The improvement of performance only comes from the characteristics of the AGEN itself without additional bells and whistles. AGEN utilize class activation maps produced by the lower layer of the network to guide segmentation branch learning spatial information and produced by the upper layer to guide segmentation branch learning semantic information, and adopt CoAM optim-

Table 4. Comparison with other end-to-end state-of-the-arts in term of mIoU(%) on the PASCAL VOC 2012 val and test sets

Dataset: val																							
Method	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
EM-Adapt [13]	67.2	29.2	17.6	28.6	22.2	29.6	47.0	44.0	44.2	14.6	35.1	24.9	41.0	34.8	41.6	32.1	24.8	37.4	24.0	68.4	54.6	33.8
CRF-RNN [14]	85.8	65.2	29.4	63.8	31.2	37.2	69.6	64.3	76.2	21.4	56.3	29.8	68.2	60.6	66.2	55.8	30.8	66.1	34.9	69.8	55.4	52.8
RRM [39]	87.9	75.9	31.7	78.3	54.6	62.2	80.5	73.7	71.2	30.5	67.4	40.9	71.8	66.2	70.3	72.6	49.0	70.7	38.4	69.7	55.1	62.6
AGEN	88.1	79.8	33.1	77.3	56.2	63.4	78.5	76.0	78.6	27.9	67.5	31.3	75.3	70.4	72.9	67.2	44.7	71.5	35.4	72.1	56.0	63.1
Dataset: test																							
Method	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
EM-Adapt [13]	76.3	37.1	21.9	41.6	26.1	38.5	50.8	44.9	48.9	16.7	40.8	29.4	47.1	45.8	54.8	28.2	30.0	44.0	29.2	34.3	46.0	39.6
CRF-RNN [14]	85.7	58.8	30.5	67.6	24.7	44.7	74.8	61.8	73.7	22.9	57.4	27.5	71.3	64.8	72.4	57.3	37.0	60.4	42.8	42.2	50.6	53.7
RRM [39]	87.8	77.5	30.8	71.7	36.0	64.2	75.3	70.4	81.7	29.3	70.4	52.0	78.6	73.8	74.4	72.1	54.2	75.2	50.6	42.0	52.5	62.9
AGEN	88.4	79.6	33.4	76.9	56.6	64.0	78.3	75.3	79.4	27.5	68.4	30.9	75.7	70.2	73.2	67.7	45.2	72.3	36.1	73.4	56.8	63.3

ize itself.

To demonstrate the efficacy and scalability of the proposed method, we extend our AGEN to a two-step approach similar to RRM. The proposed two-step method adopts segmentation results produced by end-to-end AGEN as pixel-level pseudo labels for training and evaluating the off-the-shelf fully convolutional network DeepLab-ASPP [1]. As shown in Table 5, we denote the two-step AGEN as AGEN-VGG and AGEN-ResNet based on VGG-16 [32] and ResNet-101 [40] backbones, respectively. Although some approaches used more maps for training (11k, 50k, and 970k, respectively), others used extra information such as pixel-level labels, videos, instance saliency, and saliency maps. Our two-step solution, based only on VGG-16, likewise increases performance by 0.7% and 1.4% on the val and test datasets, respectively, compared to the prior approach's most outstanding performance [26]. Additionally, we also present the comparison of our two-step approach with other methods based on ResNet. There are methods [5], [27], [29] that are based on ResNet-38 [33], which are more powerful than ResNet-101. As can be observed, the most directly comparable methods [5], [6], [27], [29], [39] are those that do not require additional training sets or supervision. Among these prior methods, reliable region mining (RRM) [39] gets the best performance similar to adopting our one-step network to produce pseudo-ground-truth. Our AGEN-ResNet achieves an mIoU value of 66.2 percent for the val set and 66.9 percent for the test set, which outperforms the RRM. Furthermore, it can be stated that our improved performance is not the result of additional training datasets or more excellent knowledge as supervision. The performance improvement is primarily due to one-step AGEN, which creates high-quality pixel-level pseudo labels to supervise training segmentation networks.

4. Qualitative results

Fig.5 shows some qualitative segmentation results achieved by our attention guided enhancement network on the PASCAL VOC val dataset. As shown in the first six columns, our end-to-end AGEN can generate accurate and comprehensive segmentation results for diverse pictures with distinct classes. We also present the segmentation results of the two-step AGEN (i.e., AGEN-VGG and AGEN-ResNet), trained with pseudo-pixel-level labels produced by the end-to-end AGEN. From the fourth and fifth rows in Fig.5, it can be seen that two-step AGEN obtain better results due to the powerful segmentation network, and demonstrate the efficacy and scalability of the proposed method. Although despite the good results, as seen in the last column of Fig.5, there are still a few typical failure instances. Interweaving objects cause the most failure cases in complex contexts are confused and misidentified.

V. Conclusions

In this paper, we present a one-step strategy for weakly supervised image semantic segmentation that relies entirely on picture-level annotations for supervision—an attention-guided enhancement network (AGEN) trained end-to-end. The AGEN contains two major modules: the Attention guided module (AGM) and the context attention module (CoAM). AGM can produce class activation maps from different layers to guide the AGEN to learn spatial and semantic information in a bottom-up fashion, while CoAM is capable of capturing the context-dependencies between class-aware feature maps produced by AGM at different stages to enhance object regions and suppress irrelevant object regions to improve the performance of segmentation further. We also extend our AGEN to a two-step approach for validating efficacy and scalability. On the PASCAL VOC 2012 semantic segmentation benchmark, experimental findings reveal that the proposed method achieves state-of-the-art performance, demonstrating the efficacy of our AGEN.

Table 5. Comparison of weakly-supervised semantic segmentation methods on the PASCAL VOC 2012 val and test sets

Method	Training	Extra information	Validation	Test
Backbone: VGG-16 [32]				
SEC [20] (ECCV'2016)	10k	–	50.7	51.7
AF-MCG [41] (CVPR'2016)	10k	Pixel-level	54.3	55.5
CrawlSeg [42] (CVPR'2017)	970k	Videos	58.1	58.7
STC [43] (TPAM'2017)	50k	Saliency	49.8	51.2
AE-PSL [3] (CVPR'2017)	10k	Saliency	55.0	55.7
DCSP [21] (BMVC'2017)	10k	Saliency	58.6	59.2
AISI [44] (ECCV'2018)	11k	Instance saliency	61.3	62.1
GAIN [22](CVPR'2018)	10k	Saliency	55.3	56.8
SeeNet [23] (NIPS'2018)	10k	Saliency	61.1	60.7
MCOF [4] (CVPR'2018)	10k	Saliency	56.2	57.6
DSRG [28] (CVPR'2018)	10k	Saliency	59.0	60.4
AffinityNet [5] (CVPR'2018)	10k	–	58.4	60.5
MDC [8] (CVPR'2018)	10k	Saliency	60.4	60.8
DSNA [6] (TMM'2019)	10k	–	55.4	56.4
FickleNet [25] (CVPR'2019)	10k	Saliency	61.2	61.8
OAA [26] (ICCV'2019)	10k	Saliency	63.1	62.8
RRM [39] (AAAI'20)	10k	–	60.7	61.0
AGEN-VGG (Ours)	10k	–	63.8	64.2
Backbone: ResNet [40]				
DCSP [20] (BMVC'2017)	10k	Saliency	60.8	61.8
AISI [44] (ECCV'2018)	10k	Instance saliency	63.6	64.5
SeeNet [23] (NIPS'2018)	10k	Saliency	63.1	62.8
MCOF [4] (CVPR'2018)	10k	Saliency	60.3	61.2
DSRG [28] (CVPR'2018)	10k	Saliency	61.4	63.2
AffinityNet* [5] (CVPR'2018)	10k	–	61.7	63.7
DSNA [6] (TMM'2019)	10k	–	58.2	60.1
FickleNet [25] (CVPR'2019)	10k	Saliency	64.9	65.3
SSDD* [29] (ICCV'2019)	10k	–	64.9	65.5
OAA [26] (ICCV'2019)	10k	Saliency	65.2	66.4
CIAN [30] (AAAI'2020)	10k	Saliency	64.3	65.3
SEAM* [27] (CVPR'2020)	10k	–	64.5	65.7
RRM [39] (AAAI'20)	10k	–	66.3	66.5
AGEN-ResNet (Ours)	10k	–	66.2	66.9

Note: * indicates methods based on ResNet-38 [10].

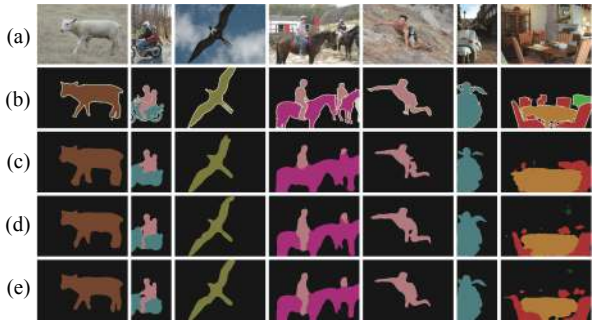


Fig. 5. Qualitative results of segmentation on the PASCAL VOC 2012 val set achieved by proposed approaches. (a) Input images; (b) Ground truth; (c) Segmentation results of end-to-end AGEN; (d) Segmentation results of two-step AGEN-VGG; (e) Segmentation results of two-step AGEN-ResNet.

References

[1] L. C. Chen, G. Papandreou, I. Kokkinos, *et al.*, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.40, no.4, pp.834–848, 2018.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, pp.234–241, 2015.

[3] Y. C. Wei, J. S. Feng, X. D. Liang, *et al.*, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp.1568–1576, 2017,.

[4] X. Wang, S. D. You, X. Li, *et al.*, "Weakly-supervised semantic segmentation by iteratively mining common object features," in *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp.1354–1362, 2018.

[5] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proceedings of 2018 IEEE/CVF Confer-*

ence on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp.4981–4990, 2018.

[6] T. Y. Zhang, G. S. Lin, J. F. Cai, *et al.*, "Decoupled spatial neural attention for weakly supervised semantic segmentation," *IEEE Transactions on Multimedia*, vol.21, no.11, pp.2930–2941, 2019.

[7] Y. Zeng, Y. Z. Zhuge , H. C. Lu, *et al.*, "Joint learning of saliency detection and weakly supervised semantic segmentation," in *Proceedings of 2019 IEEE International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), pp.7222–7232, 2019.

[8] Y. C. Wei, H. X. Xiao, H. H. Shi, *et al.*, "Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation," in *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp.7268–7277, 2018.

[9] B. L. Zhou, A. Khosla, A. Lapedriza, *et al.*, "Learning deep features for discriminative localization," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp.2921–2929, 2016.

[10] G. Papandreou, L. C. Chen, K. P. Murphy, *et al.*, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proceedings of 2015 IEEE International Conference on Computer Vision*, Santiago, Chlie, pp.1742–1750, 2015.

[11] A. Roy and S. Todorovic, "Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp.7282–7291, 2017.

[12] S. Honari, J. Yosinski, P. Vincent, *et al.*, "Recombinator networks: Learning coarse-to-fine feature aggregation," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp.5743–5752, 2016.

[13] T. Y. Lin, P. Dollár, R. Girshick, *et al.*, "Feature pyramid networks for object detection," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp.936–944, 2017.

[14] M. Everingham, L. Van Gool, C. K. I. Williams, *et al.*, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol.88, no.2, pp.303–338, 2010.

[15] D. Y. Meng and L. N. Sun, "Some new trends of deep learning research," *Chinese Journal of Electronics*, vol.28, no.6, pp.1087–1091, 2019.

[16] B. J. Zou, X. Shan, C. Z. Zhu, *et al.*, "Deep learning and its application in diabetic retinopathy screening," *Chinese Journal of Electronics*, vol.29, no.6, pp.992–1000, 2020.

[17] J. F. Dai, K. M. He, and J. Sun, "BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proceedings of 2015 IEEE International Conference on Computer Vision*, Santiago, Chlie, pp.1635–1643, 2015.

[18] D. Lin, J. F. Dai, J. Y. Jia, *et al.*, "ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp.3159–3167, 2016.

[19] A. Bearman, O. Russakovsky, V. Ferrari, *et al.*, "What's the point: Semantic segmentation with point supervision," in *Proceedings of the 14th European Conference on Computer Vision*, Amsterdam, The Netherlands, pp.549–565, 2016.

[20] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Proceedings of the 14th European Conference on Computer Vision*, Amsterdam, The Netherlands, pp.695–711, 2016.

[21] A. Chaudhry, P. K. Dokania, and P. H. S. Torr, "Discovering class-specific pixels for weakly-supervised semantic segmentation," in *Proceedings of the British Machine Vision Conference*, London, UK, pp.20.1–20.13, 2017.

[22] K. P. Li, Z. Y. Wu, K. C. Peng, *et al.*, "Tell me where to look: Guided attention inference network," in *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp.9215–9223, 2018.

[23] Q. B. Hou, P. T. Jiang, Y. C. Wei, *et al.*, "Self-erasing network for integral object attention," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montréal, Granada, pp.547–557, 2018.

[24] N. Liu and J. W. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp.678–686, 2016.

[25] J. Lee, E. Kim, S. Lee, *et al.*, "FickleNet: Weakly and semi-supervised semantic image segmentation using stochastic inference," in *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp.5267–5276, 2019.

[26] P. T. Jiang, Q. B. Hou, Y. Cao, *et al.*, "Integral object mining via online attention accumulation," in *Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*, Seoul, Korea (South), pp.2070–2079, 2019.

[27] Y. D. Wang, J. Zhang, M. N. Kan, *et al.*, Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp.12272–12281, 2020.

[28] Z. L. Huang, X. G. Wang, J. S. Wang, *et al.*, "Weakly-supervised semantic segmentation network with deep seeded region growing," in *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp.7014–7023, 2018.

[29] W. Shimoda and K. Yanai, "Self-supervised difference detection for weakly-supervised semantic segmentation," in *Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*, Seoul, Korea (South), pp.5207–5216, 2019.

[30] J. S. Fan, Z. X. Zhang, T. N. Tan, *et al.*, "CIAN: Cross-image affinity net for weakly supervised semantic segmentation," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, USA, pp.10762–10769, 2020.

[31] X. L. Zhang, Y. C. Wei, J. S. Feng, *et al.*, "Adversarial complementary learning for weakly supervised object localization," in *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp.1325–1334, 2018.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA, pp.1–14, 2015.

[33] Z. F. Wu, C. H. Shen, and A. van der Hengel, "Wider or deeper: Revisiting the ReSnet model for visual recognition," *Pattern Recognition*, vol.90, pp.119–133, 2019.

[34] X. L. Wang, R. Girshick, A. Gupta, *et al.*, "Non-local neur-

al networks," in *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp.7794–7803, 2018.

[35] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFS with Gaussian edge potentials," in *Proceedings of the 25th Annual Conference on Neural Information Processing Systems*, Granada, Spain, pp.109–117, 2011.

[36] B. Hariharan, P. Arbeláez, L. Bourdev, *et al.*, "Semantic contours from inverse detectors," in *Proceedings of 2011 International Conference on Computer Vision*, Barcelona, Spain, pp.991–998, 2011.

[37] N. Ketkar, "Introduction to pytorch," in *Deep Learning with Python*, N. Ketkar, Ed. Apress, Berkeley, CA, USA, pp.195–208, 2017.

[38] J. Deng, W. Dong, R. Socher, *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, pp.248–255, 2009.

[39] B. F. Zhang, J. M. Xiao, Y. C. Wei, *et al.*, "Reliability does matter: An end-to-end weakly supervised semantic segmentation approach," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, USA, pp.12765–12772, 2020.

[40] K. M. He, X. Y. Zhang, S. Q. Ren, *et al.*, "Deep residual learning for image recognition," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp.770–778, 2016.

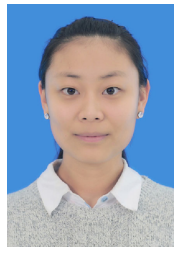
[41] X. J. Qi, Z. Z. Liu, J. P. Shi, *et al.*, "Augmented feedback in semantic segmentation under image level supervision," in *Proceedings of the 14th European Conference on Computer Vision*, Amsterdam, The Netherlands, pp.90–105, 2016.

[42] S. Hong, D. Yeo, S. Kwak, *et al.*, "Weakly supervised semantic segmentation using web-crawled videos," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp.7322–7330, 2017.

[43] Y. C. Wei, X. D. Liang, Y. P. Chen, *et al.*, "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.39, no.11, pp.2314–2320, 2017.

[44] R. C. Fan, Q. B. Hou, M. M. Cheng, *et al.*, "Associating inter-image salient instances for weakly supervised semantic segmentation," in *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, pp.371–388, 2018.

ZHANG Zhe received the M.S. degree in College of Computer Science and Technology from Jilin University, Jilin, China, in 2018, where he is currently pursuing the Ph.D. degree. He is also a member of the Key Laboratory for Symbol Computation and Knowledge Engineering of National Education Ministry, China. His main research interests include computer vision, image processing, and deep learning.
(Email: zhangzhe18@mails.jlu.edu.cn)



WANG Bilin received the B.S. degree in computer science and technology from Jilin University, Jilin, China, in 2017, where he is currently pursuing the Ph.D. degree. He is also a member of the Key Laboratory for Symbol Computation and Knowledge Engineering of National Education Ministry, China. Her main research interests include application of optimal transport and domain adaptation.
(Email: blwang19@mails.jlu.edu.cn)

YU Zhezhou (corresponding author) received the Ph.D. degree from Jilin University, in 2007. He is currently a Professor with Jilin University. He is also a member of the Key Laboratory for Symbol Computation and Knowledge Engineering of National Education Ministry, China. His research interests include computational intelligence, computer vision, image processing and embedded system application. He is a Committee Member of the Undergraduate Electronic Design Competition Organization of Jilin Province, China.
(Email: yuzz@jlu.edu.cn)

ZHAO Fengzhi received the B.E. degree from Jilin Universityin in 2017. He is currently a doctoral candidate in the computational intelligence echelon of Jilin University. His advisor is Dr. YU Zhezhou, a Professor with Jilin University. His research interests include computational intelligence, computer vision, and image segmentation.
(Email: zhaofz19@mails.jlu.edu.cn)