IEICE TRANS. FUNDAMENTALS, VOL.E107–A, NO.8 AUGUST 2024

PAPER

CyCSNet: Learning Cycle-Consistency of Semantics for Weakly-Supervised Semantic Segmentation

Zhikui DUAN[†a)], ***Member***, **Xinmei YU**[†b)], ***and*** **Yi DING**[††c)], ***Nonmembers***

SUMMARY Existing weakly-supervised segmentation approaches based on image-level annotations may focus on the most activated region in the image and tend to identify only part of the target object. Intuitively, high-level semantics among objects of the same category in different images could help to recognize corresponding activated regions of the query. In this study, a scheme called Cycle-Consistency of Semantics Network (CyCSNet) is proposed, which can enhance the activation of the potential inactive regions of the target object by utilizing the cycle-consistent semantics from images of the same category in the training set. Moreover, a Dynamic Correlation Feature Selection (DCFS) algorithm is derived to reduce the noise from pixel-wise samples of low relevance for better training. Experiments on the PASCAL VOC 2012 dataset show that the proposed CyCSNet achieves competitive results compared with state-of-the-art weakly-supervised segmentation approaches.

key words: *weakly-supervised, cycle-consistency, segmentation*

1. Introduction

Semantic segmentation is a fundamental and challenging task in computer vision, which assigns a label from a set of categories to each pixel of the image [1], [2]. Recently, convolutional neural networks (CNNs) have achieved remarkable success in semantic segmentation [3]–[5]. However, state-of-the-art semantic segmentation approaches based on CNNs require dense pixel-wise annotated data, which is prohibitively laborious. To address this issue, many weakly-supervised approaches [6], [7] have been proposed and made great progress in this area. These approaches are derived by utilizing weak annotations which can be easily obtained at low annotation costs, compared with pixel-level labels.

Due to the lack of location information in target objects, most weakly-supervised semantic segmentation approaches based on image-level supervision [6], [8] estimate the target location by generating the Class Activation Map (CAM) [9]–[11]. CAM gives high responses to discriminative parts of target objects by calculating the contribution of each region in the output according to the corresponding class. However, using only CAM for location estimation may not be optimal, because CAM mainly captures the high response region of the object, which is likely to result in incomplete boundary of objects. In order to generate a high-quality localization map, [6], [12] use localization map produced by CAM as seeds and apply region growing algorithm to expand them. On the other hand, AffinityNet [6] extracts pixel pairs relationship of an image to refine CAM under image-level annotations and achieves high performance gain for weakly-supervised semantic segmentation. However, the seeds generated from CAM are probably located outside the less discriminative parts of the object due to the lower response of these regions in CAM. As a result, existing region growing approaches cannot segment the entire object without seeds on the non-discriminative parts.

To address this problem, we propose the CyCSNet, which aims to produce more seeds from the inactivated regions of the object in CAM by learning the cycle-consistency of semantics. Instead of generating seeds by only one activation map from a single image, CyCSNet generates more seeds from the inactivated regions when common patterns of the same category exist in the training data. As shown in Fig. 1, CAM gives high responses to the main body of the ship, while CyCSNet extends the active regions and completes more key seeds from the inactivated regions (such as mast and sail in the figure) by integrating the information from other instances. The quality of final localization map can be improved after region growing due to better seeds. In a word, AffinityNet expands seeds of CAM by using the relationship between pixel pairs in the image, while CyCSNet discovers more seeds of CAM by utilizing shared semantic relationships from objects of the same category among various images.

Meanwhile, training CyCSNet collaboratively with different images is challenging. Aggregation of objects in the same category with high similarity is needed in this training. The reason is that low relevance pixel-wise samples (such as background) may produce improper gradients, which could degrade the performance of CyCSNet. To tackle this problem, we employ a scheme called Dynamic Correlation Feature Selection (DCFS), which flexibly rejects the high-level semantics feature vectors with low correlation.

The main contributions of this study are summarized as follows:

- A network called CyCSNet is proposed, which learns the semantic consistency of different images with image-level labels to improve segmentation performance.

Manuscript received June 24, 2023.
Manuscript revised October 14, 2023.
Manuscript publicized December 11, 2023.

[†]School of Electronic Information Engineering, Foshan University, Foshan, 528000, China.
[††]Hunan University of Arts and Science, Changde, 415000, China.
a) E-mail: duanzhikui@outlook.com
b) E-mail: labxmyu@fosu.edu.cn
c) E-mail: mrtbs99@gmail.com

DOI: 10.1587/transfun.2023EAP1072

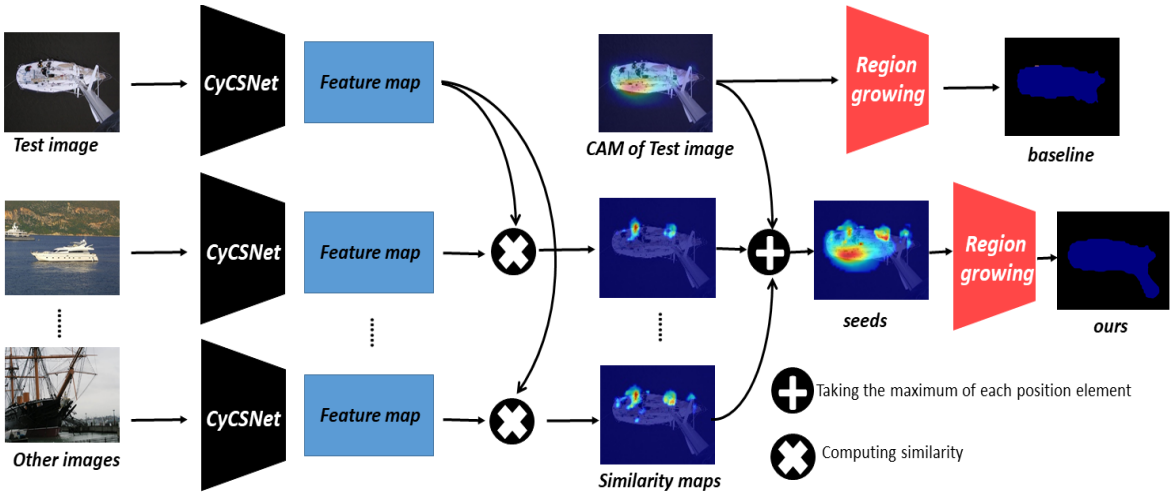

Fig. 1 Motivation: learning cycle-consistency of semantics, which uses other instances in the same category to highlight target objects and activate some inactive seeds (such as sail in the above figure). Then region growing is used to modified borderline.

- A solution named DCFS is proposed, which utilizes dynamic feature selection to reject feature vectors with low similarity of other images.
- Experiments on PASCAL VOC 2012 dataset demonstrate that the proposed scheme achieves competitive performance compared with existing weakly supervised approaches trained with only image-level annotations. These results validate that CyCSNet can improve the segmentation performance effectively by providing better seeds and activation map for region growing.

2. Related Work

2.1 Weakly-Supervised Semantic Segmentation

Most existing image semantic segmentation approaches based on image-level annotations employ class activation map (CAM) [11] for getting initial seeds of target objects. Some of them try to fix the shortcoming of CAM that only highlights the most discriminative region of objects. [13]–[15] force classification network to focus on larger areas by hiding or erasing some regions of image or feature map which are discriminative enough and have a high activation value at CAM. However, multiple computations of an image with a classification network is an expensive cost. Meanwhile, a fix hyper-parameter of times to repeat this operation could not adapt to every image.

FickleNet [8] expands CAM to cover an entire object by randomly selecting feature vectors multiple times, and then combines multiple CAMs into one. Multi-dilated convolution (MDC) [16] applies multi-channel dilated convolution blocks with various dilated rates to highlight extended regions of objects by the classifier. What's more, [12], [17], [18] use region-growing to extend the activated regions from seeds obtained by CAM. In the study by LP-CAM [19], a novel computational approach for CAM is presented. This method not only captures the usual features but also explicitly includes non-discriminative ones, ensuring that CAM provides a comprehensive coverage of entire objects. In contrast to these prior works, our approach integrates cycle-consistency regularization into weakly-supervised semantic segmentation, leading to improved results.

2.2 Cycle-Consistency

Cycle consistency between two or more samples is a successful and widely used technique in artificial intelligence. Cycle consistency has achieved great success in a lot of tasks such as image matching [20], [21], video alignment [22] and co-segmentation [23], [24]. [20] optimizes cycle consistency among feature representations of samples to improve the accuracy of dense correspondences. [22] is a self-supervised method for representation learning and aligns video well without any annotation. [22] utilize cycle consistency in videos, while our proposed scheme aims to align images, which is much difficult due to lower similarity and more elements among aligned objects. [25] employs domain adaptation through transferring source domain images to target domain style images with the help of CycleGANs. Just as [23] and [24], CyCSNet makes use of cycle consistency to conduct co-segmentation but still has some differences between ours and their models. To improve the efficiency, [23] and [24] over-segment each image in K ($K = 200$ in the original paper) super-pixels, while our method improves the efficiency by selecting vectors $V_{(x,y)}$ (red square in Fig. 2) to train CyCSNet if the value in the given position (x, y) on CAM is larger than a threshold. Background information is not only costly but also has negative effect for CyCSNet to align semantics. These two methods use GIST descriptors, and DNNs are adopted in this study. In addition, they will suffer from performance degradation if

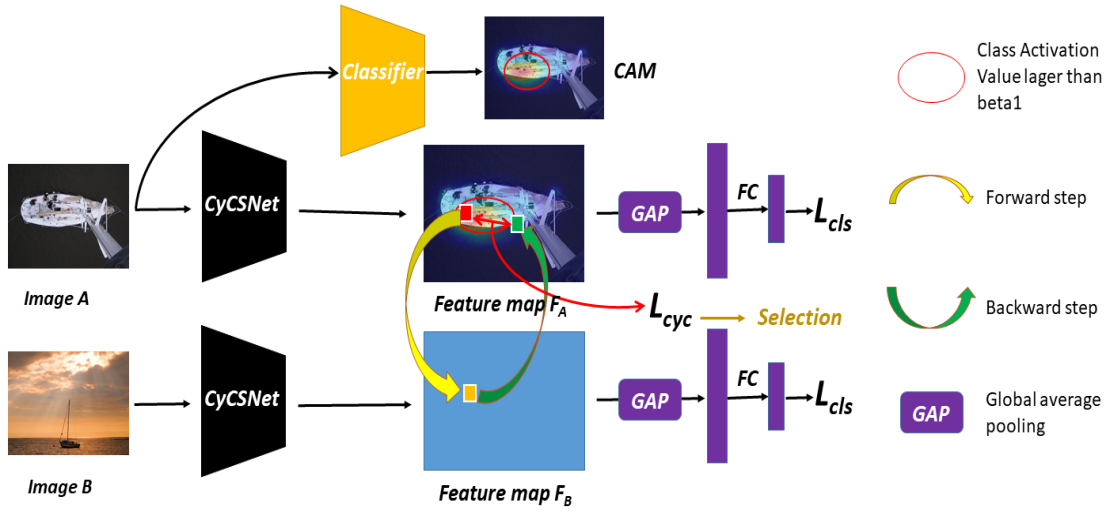

Fig. 2 Overview of training CyCSNet: input a pair of images (image A and B), and obtain feature maps by CyCSNet. Propagation step: selecting a feature vector that the activation value is larger than β_1 from feature map of A (red square in the figure) and find the nearest neighbor of the vector in feature map of B (orange square). Retrieval step: just as propagation step, we find the nearest neighbor of vector in feature maps of A (blue square). We use cycle consistency loss to make the red square closes to the blue square, and reject some feature vectors that may affect training stability. Classification loss is used to keep category information. More details can be obtained in Sect. 3.

there is a big difference between objects such as scale, while we ignore those features with low similarity degree. Most importantly, the proposed scheme is easy to combine with other weakly-supervised semantic segmentation models and advances those performances.

3. Proposed Method

3.1 Class Activation Map Construction

It is a common practice to compute class activation map (CAM) with weakly-supervised semantic segmentation approaches because of its effectiveness in locating discriminative regions. Highlighting local regions in their CAMs are usually considered as brilliant seeds of objects, which grow up to cover every object with various strategies. Here, CyCSNet utilizes CAM to pick up the foreground, which could avoid the disturbance caused by the background during training. The classification score of each specific class can be gotten through the class classification network trained with image-level labels. Then CAM can be acquired by calculating the contributions of each region based on these classification scores. It is worth mentioning that this approach requires the architecture of the classification network with global average pooling (GAP) layer followed by a classification layer. CAM can be calculated by:

$$M_c(x,y) = \sum_n ((W_n^c)^T \times f^{cam}(x,y)) \tag{1}$$

where c is the target class, and W_c denotes the weights of classification layer which connect to class c. $f^{cam}(x,y)$ represents the feature vector extracted in position (x,y), and n is the number of channels. CAM needs to be normalized so that the maximum activation value of each class is adjusted to 1.

$$M_c(x,y) \leftarrow \frac{M_c(x,y)}{\max\limits_{i,j} M_c(i,j)} \tag{2}$$

More details can be obtained in the original paper of CAM [11].

3.2 Training CyCSNet

The motivation of designing CyCSNet is to utilize multiple objects of the same class from different pictures to provide some key seeds which are not active in CAM. The information of cycle-consistency in semantics can be acquired by a pair of pictures every time. As shown in Fig. 2, the input image A and B contain at least one of the same category of their image-level labels, and then feature maps $F_A \in \mathbb{R}^{n\times h\times w}$, $F_B \in \mathbb{R}^{n\times h\times w}$ can be gotten respectively from CyCSNet. Class activation map CAM_A and CAM_B also can be gotten respectively from classification network. The training period of the proposed CyCSNet can be roughly divided into three parts: the propagation step, the retrieval step, and selection for cycle consistency.

3.2.1 Propagation Step

Motivation: As we focus on weakly-supervised setting, where we do not have traditional labeled data. Instead, we aim to use some intrinsic property of the segmentation data to design a task and create supervision. To this end, we propose a cycle-consistent based self-supervised supervision and introduce next.

Implementation: Selecting a feature vector $V_{(x,y)} \in \mathbb{R}^{1\times n}$ in F_A (There are n channels in feature maps F_A and F_B). The neighbors are defined as the similarity vectors in adjacent feature maps (F_B in this section). Then we find neighbors of $V_{(x,y)}$ from F_B and calculate the neighbor vector V_i through the neighbor feature vectors of F_B. Furthermore, we use a soft distribution instead of one-hot vectors in the propagation step because the one-hot vectors are non-differentiable. Thus, the propagation step can be described as:

$$W^f_{(x,y)} = V_{(x,y)} \times F_B \tag{3}$$

where $W^f_{(x,y)} \in \mathbb{R}^{h\times w}$ can be regarded as the similarity map of $V_{(x,y)}$ and F_B. The values of $W^f_{(x,y)}$ are usually very large so that they may out of the range of float point number after softmax, and $W^f_{(x,y)}$ is normalized as follows:

$$W^f_{(x,y)} \leftarrow softmax(\frac{\alpha \times W^f_{(x,y)}}{\max\limits_{(i,j)} W^f_{(i,j)}}) \tag{4}$$

where α is a constant hyperparameter. $h \times w$ approximately equals to 3,000, and we set α as 20. Then we calculate V_i as follows:

$$V_i = F_B \times W^f_{(x,y)} \tag{5}$$

where $V_i \in \mathbb{R}^{1\times n}$ denotes the nearest neighbor of $V_{(x,y)}$. $V_{(x,y)}$ is also the nearest neighbor of V_i, because they have the same semantics.

3.2.2 Retrieval Step

Motivation: There exist two reasons for introducing the Retrieval Step: *i*. In order to incorporate a reconstruction loss. When we translate from F_A to F_B and then back to domain F_A, we can compute a loss based on how close the cycled-back version is to the original. This reconstruction loss acts as a regularizer and helps to ensure that the learned translation preserves the content of the input. *ii*. We want to prevent from mode collapse, where the model ends up generating similar outputs for varied inputs. The cycle-consistency constraint can help in reducing the chances of mode collapse by ensuring that the distinct inputs in F_A lead to distinct translated outputs in F_B and can be cycled back correctly. By enforcing cycle-consistency, we are making sure that the model understands the inherent visual representation of paired images.

Implementation. Similar to the propagation step, during the retrieval step, we search for neighbors of V_i, which are obtained from the propagation step, from F_A. The retrieval step can be described as:

$$W^b_{(x,y)} = V_i \times F_A \tag{6}$$

where $W^b_{(x,y)} \in \mathbb{R}^{h\times w}$ is the similarity map between V_i and F_A. Then we normalize it as follow:

$$W^b_{(x,y)} \leftarrow softmax(\frac{\alpha \times W^b_{(x,y)}}{\max\limits_{i,j} W^b_{(i,j)}}) \tag{7}$$

3.2.3 Cycle Consistency and Dynamic Correlation Feature Selection

In this study, cycle consistency loss is utilized to train CyCSNet. In other words, the larger the value $W^b_{(x,y)}(x,y)$ is, the better semantic consistency it holds. However, some regions are not suitable to calculate cycle consistency loss, such as background. In that case, we use $V_{(x,y)}$ and $W^b_{(x,y)}$ to train CyCSNet, only if $CAM_{A(x,y)}$ is larger than a threshold β_1. The reason is that the smaller $CAM_{A(x,y)}$ is, the higher possibility that the region is to be the background. Another important approach to avoid negative effect caused by unsuitable samples is that rejecting $V_{(x,y)}$ if $W^b_{(x,y)}(x,y)$ is less than a threshold β_2, which means that $V_{(x,y)}$ and V_i are not sufficiently correlative. Therefore, cycle consistency loss can be written as:

$$L_{cyc} = \frac{-1}{N} \times \sum_{(x,y)\in V^{sel}} (y^{mask}_{(x,y)} \cdot \log W^b_{(x,y)}) \tag{8}$$

where V^{sel} stands for the set of co-ordinates (x,y) that $CAM_{A(x,y)}$ is larger than β_1 and $W^b_{(x,y)}(x,y)$ is larger than β_2. N denotes the number of elements which contains in the V^{sel}. We can notice that L_{cyc} is the cross entropy loss and $y^{mask}_{(x,y)}$ is the label map. Setting β_2 manually is inefficient and it is not adaptive in the training period. Thus, we propose Dynamic Correlation Feature Selection (DCFS) module that makes β_2 change with the training process to address this problem. β_2 is updated with momentum method, and β_2 is initialized to 0 at the beginning, and updated by:

$$\beta_2 \leftarrow \kappa \times \beta_2 + \frac{(1-\kappa)}{N} \sum_{(x,y)\in V^{sel}} W^b_{(x,y)}(x,y) \tag{9}$$

As for $y^{mask}_{(x,y)}$, it is optimal that the weight of the loss in the region closer to (x,y) is larger. Therefore, the expression of $y^{mask}_{(x,y)}$ can be described as follow:

$$D_{(x,y)}(i,j) = (x-i)^2 + (y-j)^2 \tag{10}$$

$$y^{mask}_{(x,y)}(i,j) = \begin{cases} r - D_{(x,y)}(i,j) & D_{(x,y)}(i,j) < r \\ 0 & D_{(x,y)}(i,j) \geqslant r \end{cases} \tag{11}$$

where r is a hyper parameter. Meanwhile, we expect CyCSNet to maintain classification performance, and multilabel soft margin loss is adopted to achieve it.

$$\begin{aligned} L_{cls} = -\sum_c (&y_c \log(\frac{1}{1+\exp(-x_c)}) + \\ &(1-y_c)\log(\frac{\exp(-x_c)}{1+\exp(-x_c)}) \end{aligned} \tag{12}$$

Algorithm 1: Learning CyCSNet

Input: training data set $\mathcal{D}$, trained classifier C
Output: CyCSNet F_θ
initialize F_θ with C
while *not converged* **do**
  Sample X_1 and X_2 from $\mathcal{D}$
  Obtain CAMs C_1 of X_1 by Eq.1;
  Calculate loss by Eq.8 and Eq.12;
  Update F_θ by Eq.13;
end
return F_θ

where c represents the prediction of each category.

Overall, the final objective of the proposed CyCSNet is delivering the optimal $\widehat{\theta_f}$ by:

$$\widehat{\theta_f} = \arg\min_{\theta_f} L_{cyc} + \lambda L_{cls} \tag{13}$$

where θ_f denotes the parameters of CyCSNet. λ is super parameters to balance the above two losses. The process of learning CyCSNet can be seen in Algorithm 1.

3.3 Class Activation Map Refinement by CyCSNet

Details of the proposed CyCSNet has already been introduced in the above sections. Next, approach on how to use CyCSNet to obtain better seeds in CAMs is presented. For an input data t, where t is an image, we can get CAM C^t by Eq. (1) and pseudo image-level label y by the classifier, and obtain a feature map with CyCSNet. Then $\mathcal{N}$ images are sampled from the training set. For each image x_s, we calculate its CAM C^s. And then select $V_{(x,y)}$ from feature map of x_s (i.e. F_s), where $C^s_{(x,y)}$ is larger than β_1. The similarity map $W^f_{(x,y)}$ can be gotten by Eq. (3) and Eq. (4). Then CAM C^x can be revised with $W^f_{(x,y)}$ by:

$$C^t_c(i,j) \leftarrow \begin{cases} \max(& C^t_c(i,j), \\ & \varepsilon W^f_{(x,y)}(i,j)) \\ & \text{if } W^f_{(x,y)}(i,j) > \frac{Bg}{\varepsilon} \\ C^t_c(i,j) & \text{otherwise} \end{cases} \tag{14}$$

where Bg denotes the preset background score, and c is the class of $W^f_{(x,y)}$ and image t must contain objects of c (i.e. $y_c = 1$).

With CyCSNet, we can obtain better seeds than CAM. However, analogous to CAM, CyCSNet can not find the complete boundary of objects. Therefore, we follow AffinityNet [6] that uses random walk step (RW) to expand seeds for better performance. In short, RW revises prediction mask with AffinityNet, which can measure the high-level semantics similarity between pixel pairs of an image and train with pseudo label CAM. More details can be found in the original paper [6].

The algorithm of the proposed CyCSNet is shown in Algorithm 2.

Algorithm 2: Using CyCSNet for testing

Input: Test image s, training data set $\mathcal{D}$,
  image nummbers $\mathcal{N}$,
  CyCSNet F_θ,trained classifier C,
  AffinityNet F_{rw};
Output: Predicted segmentation mask
iter=0;
Obtain CAMs C_s of s and class label y_s by C;
while *iter* $< \mathcal{N}$ **do**
  Samples (t, y_t) from $\mathcal{D}$;
  if $y_s \cap y_t \notin \emptyset$ **then**
    Obtain CAMs C_s by C;
    Revise mask C_s by Eq.14;
    iter $\leftarrow$ iter + 1;
  end
end
C_s=$F_{rw}(C_s)$;
return C_s

4. Experiments

4.1 Experimental Setup

(1) Dataset:

Like most weakly-supervised semantic segmentation papers, all experiments shown in this study are conducted on PASCAL VOC 2012 image segmentation benchmark [27], which contains 20 foreground object classes and one background class. Following the common practice, we train our model with the augmented training set, which totally contains 10,582 images with image-level annotations. We report the mean Intersection-over-Union (mIoU) between ground truth mask and predicted mask as the performance metric.

(2) Implementation Details:

Unless otherwise specified, we set resnet38, which strictly follows [28] as a classification network in our experiments. CyCSNet has the same structure as the classifier network and removes the last block. For faster convergence, we use the classification network as pretrain model of CyCSNet. Parameters of the entire network are optimized by stochastic gradient descent (SGD) method with the learning rate initially being set to 0.01 and halved every epoch. For all experiments, the background score is set to 0.2, and the ε in Eq. (14) is set to 2.0.

The scale parameter α in Eq. (4) and Eq. (7) is set to 20, and activation threshold β_1 (see in Sect. 3.2.3) is 0.3. Momentum weight κ in Eq. (9) is 0.95 by default. Meanwhile, we set the super parameter λ in Eq. (13) to 0.1 and r in Eq. (11) to 5 in all the following experiments. Furthermore, unless otherwise stated, the backbone of our method is as same as the baseline.

4.2 Comparison with the State of the Art

(1) Results on PASCAL VOC 2012 Training Set:

Self-supervised scale equivariant network (SEENet) [26] is a

Table 1 Quantitative results of the proposed approach and baselines on the on the PASCAL VOC 2012 training set. Cyc can be viewed as CyCSNet.

Mothod	RW [6]	Cyc	mIoU	comparison
CAM [11]			47.7	-
CAM		✓	47.9	+0.2
CAM	✓		61.4	-
CAM	✓	✓	63.0	+1.6
SEENet [26]			49.8	-
SEENet		✓	49.9	+0.1
SEENet	✓		62.1	-
SEENet	✓	✓	63.1	+1.0

Table 2 Quantitative results of the proposed approach and baselines on the on the PASCAL VOC 2012 validation set. RW is random walk with AffinityNet. Cyc means CyCSNet. * indicates that additional data is being used.

Mothod	Supervision	Saliency	mIoU
What's Point [29]	Point	-	46.0
RAWK [30]	Scribble	-	61.4
ScribbleSup [31]	Scribble	-	63.1
BoxSub [32]	Bbox	-	62.0
WSSL [33]	Bbox	-	62.6
SDI [7]	Bbox	-	65.7
STC [34]	Image-level	✓	49.8
AdvErasing [35]	Image-level	✓	55.0
SeeNet [36]	Image-level	✓	63.1
DSRG [12]	Image-level	✓	61.4
FickleNet [8]	Image-level	✓	64.9
DeepLab [5]	Full	-	67.6
ResNet38 [28]	Full	-	80.8
EM-Adapt [33]	Image-level	-	38.2
MIL [37]	Image-level	-	42.0
SEC [17]	Image-level	-	50.7
TransferNet* [38]	Image-level	-	52.1
Saliency* [39]	Image-level	-	55.7
MCNN* [40]	Image-level	-	38.1
CrawlSeg* [41]	Image-level	-	58.1
AffinityNet [6]	Image-level	-	61.7
AffinityNet+LPCAM	Image-level	-	63.0
AffinityNet+Cyc(ours)	Image-level	-	62.7
AffinityNet+LPCAM+Cyc(Ours)	Image-level	-	**63.5**

method that utilizes scale equivariance as supervision information to improve the performance of weakly-supervised semantic segmentation. Backbone networks of all methods are resnet38, and the number of images $\mathcal{N}$ mentioned in Sect. 3.3 is set to 64 if the methods use random walk step (RW). And $\mathcal{N}$ equals to 16 if methods perform without RW. For fair comparison, experiments are conducted in training set as in [26]. The quantitative results are shown in Table 1. For all scenarios, the integration of basic model and CyCSNet has better performance than basic model. Without using region growing (RW in this experiment), the proposed approach only slightly outperforms baselines on mIoU. The reason is that CyCSNet activates more seeds (as shown in Fig. 1) but they can not cover the entire regions of objects. Combining with RW, CyCSNet can bring 1.6 mIoU improvement for CAMs+RW, and 1.0 mIoU improvement for SSENet+RW. And our method achieves state-of-the-art performance under the same setting of image-level weak supervision.

Table 3 Quantitative results of the proposed approach and baselines on the PASCAL VOC 2012 validation set and the backbone is set to VGG16.

Mothod	Training	mIoU
Supervision: Image-level and additional annotations		
MIL-seg [37]	700K	42.0
STC [34]	50K	49.8
TransferNet [38]	70K	52.1
CrawlSeg [41]	970K	58.1
AISI [42]	11K	61.3
Supervision: Image-level annotations only		
SEC [17]	10K	50.7
CBTS-cues [43]	10K	52.8
TPL [44]	10K	53.1
AE_PSL [35]	10K	55.0
DCSP [45]	10K	58.6
GAIN [13]	10K	55.3
MCOF [46]	10K	56.2
DSRG [12]	10K	59.0
MDC [16]	10K	60.4
AffinityNet [6]	10K	58.8
AffinityNet+Cyc(ours)	10K	**60.5**

Table 4 Quantitative results of the proposed approach on the PASCAL VOC 2012 validation set when $\mathcal{N}$ takes different values.

$\mathcal{N}$	mIoU	comparison
AffinityNet($\mathcal{N}$=0)	61.7	-
$\mathcal{N}$=8	61.9	+0.2
$\mathcal{N}$=16	62.0	+0.3
$\mathcal{N}$=32	62.2	+0.5
$\mathcal{N}$=64	62.5	+0.8
$\mathcal{N}$=128	**62.7**	+1.0
$\mathcal{N}$=256	62.7	+1.0

(2) Results on PASCAL VOC 2012 Validation Set:

The number of images $\mathcal{N}$ is set to 128. The backbone of classification network is resnet38 in this experiment. The quantitative results on validation set images are shown in Table 2. AffinityNet [6] achieves a very successful result with only image-level supervision, and the source code is available on GitHub platform. Thus, we choose AffinityNet as baseline. As a result, with the help of CyCSNet, we can improve the AfinityNet with 1.0% mIoU performance on the very high baseline. Specifically, the proposed approach recovers 77.6% of its upper bound (i.e. trained with full pixel-level annotations) when `resnet38` is used as the backbone.

Furthermore, the proposed model can enhance State-of-the-Art Activation Map Extraction. We augment our CYCSNet with the state-of-the-art LPCAM method [19]. It becomes evident that the combined deployment of LPCAM and CYCSNet yields superior results in comparison to using either model individually. Specifically, this combined approach outperforms CYCSNet in isolation by 0.8%. It's worth noting that, in our experiments, we observed that LPCAM requires a longer convergence time compared to CYCNet alone. To maintain a fair comparison, we streamlined this process and conducted experiments under identical settings. These results underscore the synergistic performance improvement achieved by these two models when used together. It becomes apparent that CYCSNet represents an effective means to enhance the performance of weakly-

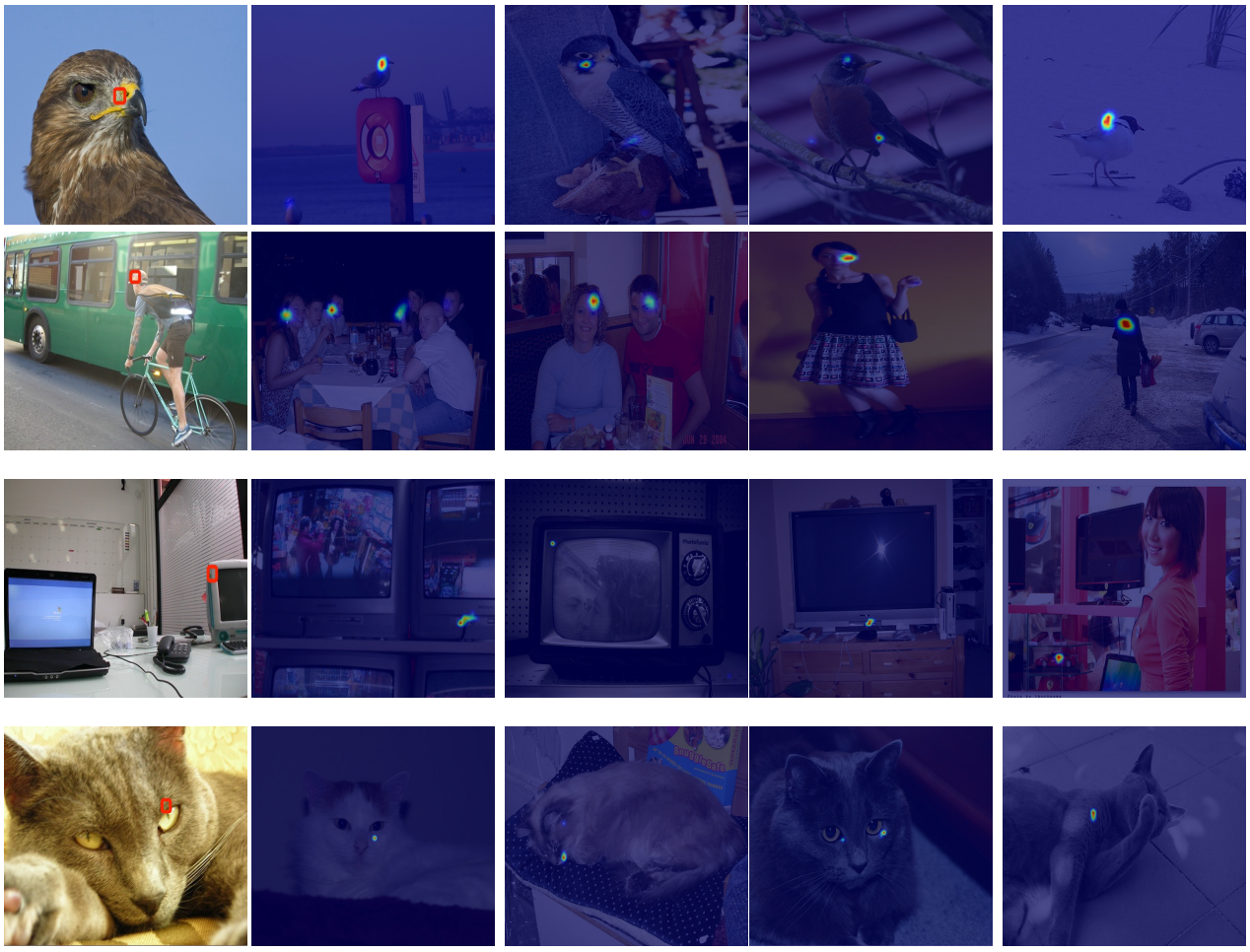

Fig. 3 Visualization of CyCSNet: Images in the first column are source images, and red points in images are vectors to map i.e. $V_{(x,y)}$ in Eq. (3). Images shown from the second to fifth columns are the similarity map (Eq. (4)) between $V_{(x,y)}$ and other images. It is worth mentioning that images in the last column are the failure cases.

supervised semantic segmentation.

(3) Results of Another backbone:

When the backbone is set to VGG16, the experimental results are shown in Table 3. The proposed method achieves well performance with VGG16 as a classification network. The result of our method does not obviously outperform MDC [16], but the most important thing is that CyCSNet can make the state of the art model, such as AffinityNet, to be more powerful.

4.3 Effects of $\mathcal{N}$

In the following, we will explore the impact of super parameter $\mathcal{N}$ (see in Sect. 3.3). In this experiment, we use resnet38 as the backbone and report the results on the PASCAL VOC 2012 validation set. Except for super parameter $\mathcal{N}$, all settings are the same as that mentioned in Sect. (2).

The results of this experiment are shown in Table 4. At the beginning, the result of mIoU increases with the increase of $\mathcal{N}$. However, when $\mathcal{N} = 128$, the result of the network just slightly outperforms than that $\mathcal{N} = 256$. The reason is that more pictures will significantly activate more seeds of different parts of objects. However, too much pictures may incorrectly activate some seeds.

4.4 Analysis of Dynamic β_2

In this section, we conduct a comprehensive analysis of the influence of β_2 as outlined in Eq. (9). Our initial approach involves setting β_2 to a constant value, establishing it as the baseline for subsequent experiments. We then employ our Dynamic Correlation Feature Selection (DCFS) model and present the results in Table 5.

Our findings indicate that dynamic updates outperform pre-defined values, and dynamic updates with moment values yield the best results.

To provide a more detailed understanding of the evolving behavior of β_2 over time, we chart its progression across training epochs in Fig. 4. An in-depth analysis of this chart reveals a distinct pattern as the value gradually converges to 0.4, and the loss function stabilizes without further reduc-

Table 5 The Dynamic β_2 Ablation Study: In this study, we examine the impact of varying coefficients, with the first row representing the baseline.

Method	β_2	mIoU
AffinityNet	-	61.7
AffinityNet	0.1	62.1
AffinityNet	0.3	62.2
AffinityNet	1.0	61.8
AffinityNet	dynamic wo moment	62.4
AffinityNet	dynamic	**62.7**

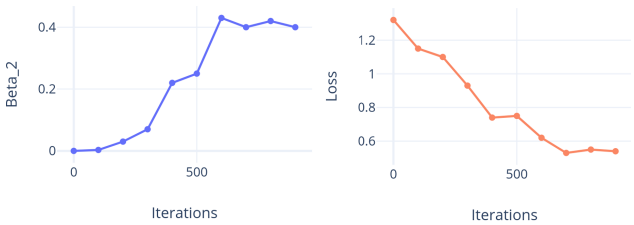

Fig. 4 The evolution of β_2 over time: we present the average β_2 values and observe a smooth transition, attributed to the introduction of moment values. These values ultimately converge to 0.4.

tion.

4.5 Visualization of CyCSNet

To better illustrate the effectiveness of CyCSNet, we conduct a visual experiment. Some samples of the results are shown in Fig. 3. Even if there are some differences between image pairs, such as scale sizes, posture, object numbers, CyCSNet can calculate a good similarity map of high-level semantics.

4.5.1 Analysis of Failure Cases

In this experimental section, we present an analysis of failure cases encountered with CyCSNet. As illustrated in Fig. 3, we display both the source images and their corresponding map vectors in the target frame. It is apparent that on occasion, the region of focus in the map vector may exhibit similarity to another region, as seen in the second row. Moreover, there are instances when the method erroneously directs attention to irrelevant regions, such as in the last row, where the attention is drawn to the cat's body instead of its eye. These issues can be attributed, in part, to the high similarity between regions in the source and target images.

5. Conclusion

This paper has introduced a novel CyCSNet network for weakly supervised semantic segmentation based on image-level annotations. Experiments on PASCAL VOC 2012 demonstrate that the proposed CyCSNet outperforms state-of-the-art approaches. Since the proposed method is trained by utilizing other images information to generate better seeds than CAM, it can achieve much better performance after region growing. To train CyCSNet better, we proposed DCFS (Dynamic Correlation Feature Selection) to remove feature vectors which may not be beneficial for learning CyCSNet.

Acknowledgments

This work was supported by the Guangdong Basic and Applied Basic Research Foundation (No.2019A1515110127 and No.2021B1515120025, No.2020A1515111107), the key Laboratory in the higher education institutions of educational commission of guangdong province, China (Grant No.2021KSYS008).

References

[1] T. Yang, Y. Yoshimura, A. Morita, T. Namiki, and T. Nakaguchi, "Pyramid predictive attention network for medical image segmentation," IEICE Trans. Fundamentals, vol.E102-A, no.9, pp.1225–1234, Setp. 2019.
[2] M.D. Sulistiyo, Y. Kawanishi, D. Deguchi, I. Ide, T. Hirayama, J.Y. Zheng, and H. Murase, "Attribute-aware loss function for accurate semantic segmentation considering the pedestrian orientations," IEICE Trans. Fundamentals, vol.E103-A, no.1, pp.231–242, Jan. 2020.
[3] G. Bertasius, L. Torresani, S.X. Yu, and J. Shi, "Convolutional random walk networks for semantic image segmentation," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.6137–6145, 2017.
[4] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A.L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," International Conference on Learning Representations (ICLR), 2015.
[5] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A.L. Yuille, "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, 2018.
[6] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4981–4990, 2018.
[7] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," Conference on Computer Vision and Pattern Recognition (CVPR), pp.1665–1674, 2017.
[8] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon, "FickleNet: Weakly and semi-supervised semantic image segmentation using stochastic inference," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.5267–5276, 2019.
[9] A. Chattopadhay, A. Sarkar, P. Howlader, and V.N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp.839–847, 2018.
[10] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," Proc. IEEE International Conference on Computer Vision, pp.618–626, 2017.
[11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2921–2929, 2016.
[12] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.7014–7023, 2018.
[13] K. Li, Z. Wu, K.C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.9215–9223, 2018.
[14] K.K. Singh and Y.J. Lee, "Hide-and-seek: Forcing a network to

be meticulous for weakly-supervised object and action localization," 2017 IEEE International Conference on Computer Vision (ICCV), pp.3544–3553, 2017.
[15] 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," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1568–1576, 2017.
[16] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T.S. Huang, "Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.7268–7277, 2018.
[17] A. Kolesnikov and C.H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," European Conference on Computer Vision (ECCV), pp.695–711, 2016.
[18] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," Advances in Neural Information Processing Systems, pp.109–117, 2011.
[19] Z. Chen and Q. Sun, "Extracting class activation maps from non-discriminative features as well," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.3135–3144, 2023.
[20] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A.A. Efros, "Learning dense correspondence via 3D-guided cycle consistency," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.117–126, 2016.
[21] T. Zhou, Y. Jae Lee, S.X. Yu, and A.A. Efros, "FlowWeb: Joint image set alignment by weaving consistent, pixel-wise correspondences," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.1191–1200, 2015.
[22] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "Temporal cycle-consistency learning," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.1801–1810, 2019.
[23] F. Wang, Q. Huang, and L.J. Guibas, "Image co-segmentation via consistent functional maps," Proc. IEEE International Conference on Computer Vision, pp.849–856, 2013.
[24] F. Wang, Q. Huang, M. Ovsjanikov, and L.J. Guibas, "Unsupervised multi-class joint image segmentation," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.3142–3149, 2014.
[25] J. Hoffman, E. Tzeng, T. Park, J.Y. Zhu, P. Isola, K. Saenko, A.A. Efros, and T. Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," arXiv preprint arXiv:1711.03213, 2017.
[26] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised scale equivariant network for weakly supervised semantic segmentation," arXiv preprint arXiv:1909.03714, 2019.
[27] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," Int. J. Comput. Vis., vol.88, no.2, pp.303–338, 2010.
[28] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the ResNet model for visual recognition," Pattern Recognition, vol.90, pp.119–133, 2019.
[29] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," European Conference on Computer Vision, pp.549–565, Springer, 2016.
[30] P. Vernaza and M. Chandraker, "Learning random-walk label propagation for weakly-supervised semantic segmentation," Conference on Computer Vision and Pattern Recognition (CVPR), pp.2953–2961, 2017.
[31] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation," Conference on Computer Vision and Pattern Recognition (CVPR), pp.3159–3167, 2016.
[32] J. Dai, K. He, and J. Sun, "BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," IEEE International Conference on Computer Vision (ICCV), pp.1635–1643, 2015.
[33] G. Papandreou, L.C. Chen, K.P. Murphy, and A.L. Yuille, "Weakly- and semi-supervised learning of a DCNN for semantic image segmentation," IEEE International Conference on Computer Vision (ICCV), pp.1742–1750, 2015.
[34] Y. Wei, X. Liang, Y. Chen, X. Shen, M.M. Cheng, J. Feng, Y. Zhao, and S. Yan, "STC: A simple to complex framework for weakly-supervised semantic segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol.39, no.11, pp.2314–2320, 2016.
[35] 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," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.6488–6496, 2017.
[36] Q. Hou, P. Jiang, Y. Wei, and M.M. Cheng, "Self-erasing network for integral object attention," Advances in Neural Information Processing Systems, pp.549–559, 2018.
[37] P.O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1713–1721, 2015.
[38] S. Hong, J. Oh, H. Lee, and B. Han, "Learning transferrable knowledge for semantic segmentation with deep convolutional neural network," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.3204–3212, 2016.
[39] S.J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele, "Exploiting saliency for object segmentation from image level labels," Conference on Computer Vision and Pattern Recognition (CVPR), pp.5038–5047, 2017.
[40] P. Tokmakov, K. Alahari, and C. Schmid, "Weakly-supervised semantic segmentation using motion cues," European Conference on Computer Vision, pp.388–404, 2016.
[41] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han, "Weakly supervised semantic segmentation using web-crawled videos," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2224–2232, 2017.
[42] R. Fan, Q. Hou, M.M. Cheng, G. Yu, R.R. Martin, and S.M. Hu, "Associating inter-image salient instances for weakly supervised semantic segmentation," Proc. European Conference on Computer Vision (ECCV), pp.367–383, 2018.
[43] A. Roy and S. Todorovic, "Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.3529–3538, 2017.
[44] D. Kim, D. Cho, D. Yoo, and I. So Kweon, "Two-phase learning for weakly supervised object localization," Proc. IEEE International Conference on Computer Vision, pp.3534–3543, 2017.
[45] A. Chaudhry, P.K. Dokania, and P.H. Torr, "Discovering class-specific pixels for weakly-supervised semantic segmentation," arXiv preprint arXiv:1707.05821, 2017.
[46] X. Wang, S. You, X. Li, and H. Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," IEEE Conference on Computer Vision and Pattern Recognition, pp.1354–1362, 2018.

Zhikui Duan was born in Nei Mongolia, China, in 1985. He received the B.S. degree from Central South University, Changsha, China, in 2008, the M.S. degree from National University of Defense Technology, Changsha, China, in 2011 and Ph.D. degree from Sun Yat-sen University, Guangzhou, China, in 2015. Since 2016, he has worked in Foshan University. His research interests include audio and video processing, analog integrated circuits.

Xinmei Yu was born in Nei Mongolia, China, in 1974. She received the B.S. and M.S. degrees in Guilin University of Electronic Technology. She is currently an associate professor with the school of electronic information engineering, Foshan University. Her research interests include electronic circuit and system, audio and video processing.

Yi Ding received the B.S. and M.S. degrees in National University of Defense Technology, in 2000 and 2004, respectively, and Ph.D. degree in communication and information systems, School of Information Science and Technology, Sun Yat-sen University, Guangzhou, China, in 2014. He is working in Hunan University of Arts and Science. His research interests include audio and video processing, Internet of things and artificial intelligence.