Harmonic Feature Activation for Few-Shot Semantic Segmentation

Binghao Liu, Jianbin Jiao, *Member, IEEE*, and Qixiang Ye, *Senior Member, IEEE*

***Abstract*— Few-shot semantic segmentation remains an open problem because limited support (training) images are insufficient to represent the diverse semantics within target categories. Conventional methods typically model a target category solely using information from the support image(s), resulting in incomplete semantic activation. In this paper, we propose a novel few-shot segmentation approach, termed harmonic feature activation (HFA), with the aim to implement dense support-to-query semantic transform by incorporating the features of both query and support images. HFA is formulated as a bilinear model, which takes charge of the pixel-wise dense correlation (bilinear feature activation) between query and support images in a systematic way. HFA incorporates a low-rank decomposition procedure, which speeds up bilinear feature activation with negligible performance cost. In addition, a semantic diffusion procedure is fused with HFA, which further improves the global harmony and local consistency of the feature activation. Extensive experiments on commonly used datasets (PASCAL VOC and MS COCO) show that HFA improves the state-of-the-arts with significant margins. Code is available at https://github.com/Bibikiller/HFA.**

***Index Terms*— Few-shot learning, semantic segmentation, harmonic activation, semantic diffusion, bilinear model.**

I. Introduction

THANKS to the large-scale datasets with dense annotation, Convolutional Neural Networks (CNNs) have made unprecedented progress in computer vision tasks [1]–[5]. Nevertheless, large-scale dense annotation usually requires great human effort and time cost, while models trained on such datasets often fail to handle novel object categories. As a promising direction, few-shot learning, *e.g.*, few-shot semantic segmentation, has been proposed to solve those problems. It targets at learning high-performance models upon training samples while generalizing the model to novel categories with only a few support images.

In the deep learning era, few-shot learning is often exploited in the metric learning framework with a two-stream structure consisting of a support branch and a query branch [6]–[12]. The support branch aims to extract specific semantic representation from support image(s), and uses such representation to guide the query branch for semantic segmentation. However, many existing works solely leverage the semantic representation from support image(s), *i.e.*, extracting the foreground features of support image(s) and concatenating the extracted features for linear activation. Such linear activation based on feature concatenation leads to insufficient information interaction between support and query images, Fig. 1(a). Given limited support images, it is hard to learn complete semantics for objects of various scales, perspectives, and poses. This consequently causes the incomplete semantic activation and false/missing segmentation.

In this paper, we propose the harmonic feature activation (HFA) approach to transform semantics from support to query images while considering the intra-image semantic consistency. Semantic activation is formulated as a bilinear model, which takes charge of the pixel-wise dense correlation (bilinear feature activation) between query and support images. This is implemented as a tensor operation between support and query features. HFA thus leverages complete features of support image(s), instead of pooled features, for dense semantic interaction, Fig. 1. When the dimensionality of support and query features is high, however, the scale of the tensor operation is large and the efficiency is low. We thus further propose a low-rank decomposition procedure, which speeds up bilinear feature activation by decomposing the tensor to three matrices and a small core tensor. To guarantee the semantic consistency within the query image, HFA further incorporates a semantic diffusion module. For the variation of object scales, perspectives and poses, semantic activation of the query image would be incomplete, *i.e.*, when some object parts are well activated, others could be missed, Fig. 1(a). The semantic diffusion module, as a complementary to support-to-query activation, is used to refine object extent by defining an iterative pixel-level semantic propagation. With semantic diffusion, the under-activated features in object parts are refined and the semantic consistency of target objects is enhanced, Fig. 1(b).

The contributions of this work are summarized as follows:

- We propose the Harmonic Feature Activation (HFA) approach, defining a systematic way for support-to-query semantic transform based on bilinear feature activation and semantic diffusion modules.

Manuscript received July 31, 2020; revised December 22, 2020, January 13, 2021, and January 19, 2021; accepted February 6, 2021. Date of publication February 17, 2021; date of current version February 26, 2021. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61836012 and Grant 61771447, in part by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDA27000000, and in part by CAAI-Huawei MindSpore Open Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yi Yang. *(Corresponding author: Qixiang Ye.)*

The authors are with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: liubinghao18@mails.ucas.ac.cn; jiaojb@ucas.ac.cn; qxye@ucas.ac.cn).

Digital Object Identifier 10.1109/TIP.2021.3058512

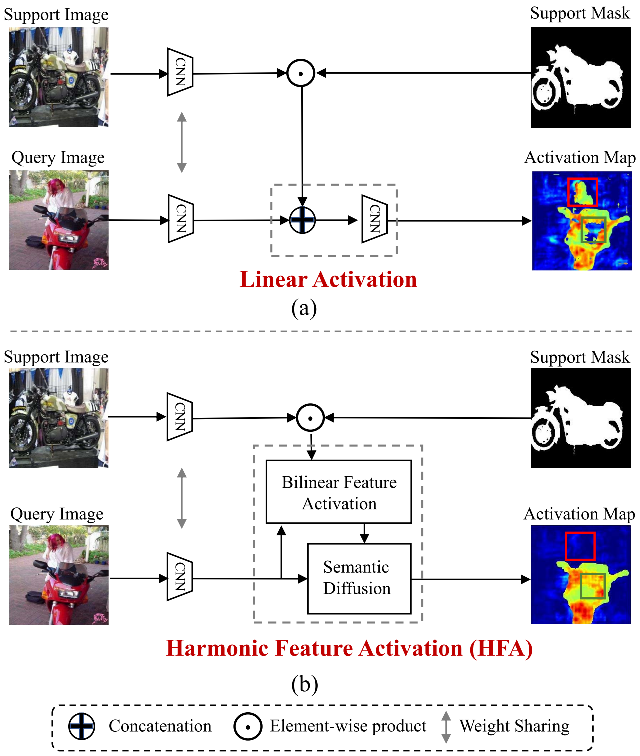


Fig. 1. Comparison of harmonic feature activation (HFA) with linear feature activation. Due to the scale, perspective, and pose variation of objects, there is a semantic gap between objects in support image(s) and query images. Linear feature activation represents a target category solely using information from the support image(s) while unfortunately ignoring the semantics within the query image, which causes incomplete activation (green box) and false activation (red box). The proposed HFA leverages bilinear feature activation and semantic diffusion to fuse the information from both query and support images for harmonic semantic activation.

- We propose a low-rank decomposition strategy to approximate the bilinear feature activation, providing an efficient way for dense semantic transform.
- We achieve new state-of-the-art performance on the PASCAL VOC and MS COCO semantic segmentation datasets. Particularly, on the large-scale MS COCO dataset, we improve the 1-shot segmentation performance by 3.81%, which is a significant margin.

II. Related Work

A. Semantic Segmentation

Various supervised and weakly supervised segmentation methods are based on the fully convolution network. DeepLab [13] adopts atrous spatial pyramid pooling (ASPP) to explicitly control the resolution at which feature responses are computed within Deep Convolutional Neural Networks. The weakly supervised segmentation approach [14] uses the recursive semantic segmentation framework based on image-level category labels. Ontology-based semantic image segmentation (OBSIS) [15] jointly models image segmentation and object detection. Relevant researches about semantic segmentation have provided fundamental techniques, *e.g.*, multi-scale feature aggregation [2] and ASPP [13] for few-shot semantic segmentation.

B. Few-Shot Learning

State-of-the-art methods for few-shot learning can be roughly categorized as either: metric learning [16]–[20], meta-learning [21]–[24], or data augmentation [25], [26]. Metric learning methods measure the distances between images/regions. Meta-learning based approaches improve optimization strategies or loss functions, which speed up learning and updating of parameters with few examples from novel categories. Data augmentation methods typically generate new samples for unseen categories [25], [26].

Prototypical models [8], [9], [27] that convert the spatial semantic information of objects to the convolutional channels have achieved the state-of-the-art results on few-shot learning. While these methods leverage the semantic information in support image(s), few of them consider incorporating the information from query images. Prototype-relation Network [28] leverages prototypes and semantic relations for few-shot recognition, presenting a new loss function which takes both inter-class and intra-class distances into account. The reinforcement learning method [29] leverages sampling to de-correlate the semantics within an image, and extracts varying sequences of patches on every forward-pass with discriminative information observed. This can be viewed as a form of "learned" data augmentation in the sense searching for different sequences of patches within an image and performs classification with the aggregation of the extracted features, resulting in improved performance. The semantic selection method [30] pursues a universal representation by training a set of semantically different feature extractors. It then uses the universal representation to automatically select the most relevant representation for semantic activation.

C. Few-Shot Segmentation

Early few-shot segmentation methods usually adopt a parametric module, which fuses information extracted from the support image(s) to segment the query image with a few convolutional operations. In [31] support features are concatenated with query ones to activate features within object regions. The activated features are fed to the following convolutional layers to generate segmentation masks. In [7], masked average pooling is utilized to extract foreground/background information within support image(s). CANet [8] consists of a two-branch model which performs feature matching between the support image(s) and the query image. It also proposes an iterative refinement module which iteratively refines the segmentation results. FWB [10] focuses on improving the representative capability of support features by leveraging foreground-background feature differences. Based on features learned on support-query image pairs, the target object in the query image is segmented by using a metric-based comparison between the class feature vector and the query feature maps.

PANet [9] offers high-quality prototypes that are representative for each semantic class and meanwhile discriminative for different classes. During training, it introduces a special strategy to perform data argumentation by exchanging the roles of support and query images. Considering the semantic gap between support and query images, however, the model

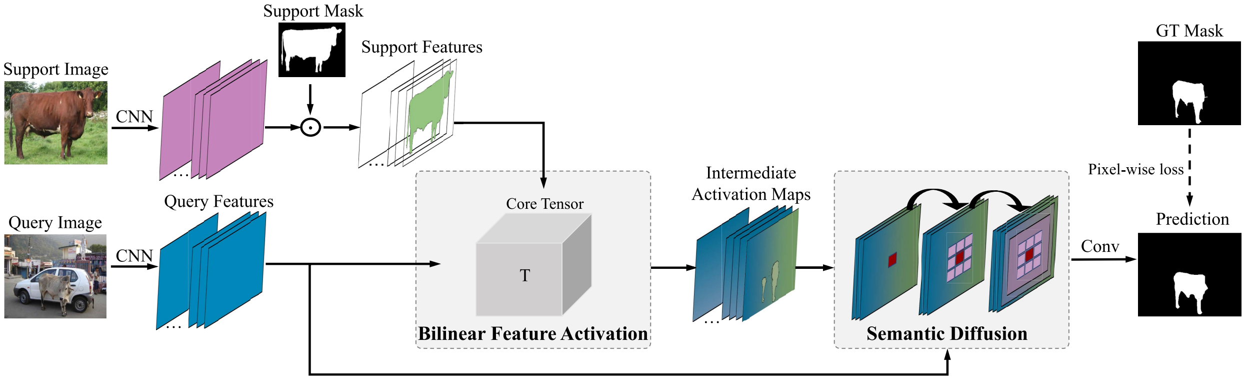


Fig. 2. Flowchart of the proposed harmonic feature activation (HFA), which consists of a bilinear feature activation module and a semantic diffusion module.

is challenged by object categories of similar semantics. PGNet [11] uses attentive graph reasoning to propagate label information from support image(s) to the query image. The graph attention mechanism establishes the element-to-element correspondence across structured data by learning attention weights between nodes. Nevertheless, when transferring semantics from support to query image, the semantic consistency within the query image is ignored. The cross-reference network [12] concurrently makes predictions for both the support image(s) and the query image. With a cross-reference mechanism, it finds the co-occurrent objects, thereby improving the semantic transfer between support and query images. However, the cross-reference network remains using a single prototype to perform segmentation, which increases semantic ambiguity and misses object parts.

D. Bilinear Model

In this study, we aim to transfer dense semantics from support to query images by a bilinear model, which strengthens the response of target category features in the way of bilinear pooling. Existing works [32] have effectively reduced the descriptor dimension of bilinear pooling by performing a Random Projection, making it possible to compute the high-dimensional descriptor without using explicit tensor operation. Lin et al [33] elaborate the Bilinear pooling with CNNs. They find that the matrix square-root normalization outperforms alternative schemes such as the matrix logarithm normalization when combining element-wise square-root with L_2 normalization. We use the bilinear model to activate and fuse features in the few-shot setting. The purpose is to strengthen the response of target category features in a preciser way. We also introduce a low-rank decomposition procedure, which speeds up the bilinear model with plausible approximation.

III. Methodology

We first formulate semantic activation as a bilinear model, which fuses the query and support features together and activates the features related to target object categories. We then introduce low-rank tensor decomposition to approximate the bilinear model and improve computational efficiency. Finally, we propose the semantic diffusion module, which propagates confidence among query features leveraging the semantic consistency to improve feature activation. The semantic segmentation mask is obtained after a few convolutional operations on the activated feature maps.

A. Bilinear Model

The aim of few-shot semantic segmentation is to classify each pixel within object extent to a pre-defined category as well as classifying other pixels to the background. The segmentation model requires to fully leverage the limited semantics in few-shot support image(s) and the query image to activate full object extent in the query image. To fulfill this purpose, the few-shot segmentation problem is formulated as a bilinear model [34], which leverages the pair-wise correlation between support and query features to activate query features. A general bilinear model is defined as

$$f(X,Y) = T \times_1 X \times_2 Y, \tag{1}$$

where X and Y respectively denote the matrices to be fused. T denotes a core tensor which fuses the input matrices using a bilinear model. $f(X,Y)$ denotes the output of the bilinear model. $\times_i$ represents the i-mode product [35] between a tensor and a matrix.

For few-shot segmentation, Fig. 2, we first extract features from the support and query images using a CNN. We then resize the support mask and multiply it with support features in a pixel-wised manner to highlight features corresponding to objects of interests and depress those corresponding to background regions. The query features and the highlighted support features are denoted as $Q_f \in \mathbb{R}^{D_q \times H \times W}$ and $S_f \in \mathbb{R}^{D_s \times H \times W}$. D_q and D_s denote the numbers of feature channels. H and W denote the height and width of the maps. Q_f and Q_s are respectively reshaped to $Q \in \mathbb{R}^{HW \times D_q}$ and $S \in \mathbb{R}^{HW \times D_s}$. The bilinear model is then applied to correlate the semantics of support and query features to activate the query features, as

$$A = T \times_1 S \times_2 Q, \tag{2}$$

where $A \in \mathbb{R}^{HW \times HW \times D_o}$ denotes the activated features. D_o denotes the feature channel number of A. $T \in \mathbb{R}^{D_s \times D_q \times D_o}$ denotes a core tensor which fuses S and Q in a pair-wise fashion.

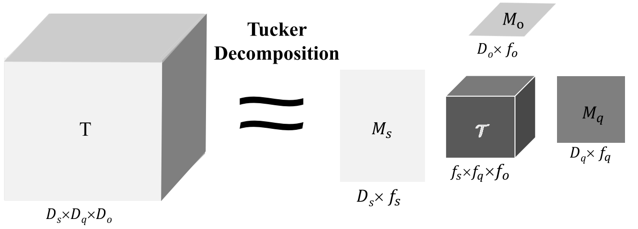


Fig. 3. Tucker decomposition of the core tensor.

B. Low-Rank Decomposition

When the dimensionality of support and query features is high, the size of core tensor T is very large, which means high computational and memory cost. For example, for $D_s = D_q = 256$, $D_o = 256$, T has 16777216 parameters. We thereby introduce the low-rank decomposition to approximate the operation defined in Eq. 2 while significantly reducing the computational cost.

As shown in Fig. 3, we use the tucker decomposition [35] to decompose a large tensor to three matrices and a small core tensor. For a 3rd-order tensor T, the tucker decomposition is defined as

$$T = \mathcal{T} \times_1 M_s \times_2 M_q \times_3 M_o, \tag{3}$$

which converts $T \in \mathbb{R}^{D_s \times D_q \times D_o}$ to three unitary 2D matrices, $M_s \in \mathbb{R}^{D_s \times f_s}$, $M_q \in \mathbb{R}^{D_q \times f_q}$ and $M_o \in \mathbb{R}^{D_o \times f_o}$, and a smaller core tensor, $\mathcal{T} \in \mathbb{R}^{f_s \times f_q \times f_o}$. f_s, f_q and f_o are the dimensions of $\mathcal{T}$, which are usually smaller than D_s, D_q and D_o.

Accordingly, Eq. 2 is re-written as

$$A = \mathcal{T} \times_1 (S \times M_s) \times_2 (Q \times M_q) \times_3 M_o, \tag{4}$$

where M_s and M_q are 2-D matrices which project the support and query features (S and Q) into an embedding space and the features are denoted as $\hat{S} \in \mathbb{R}^{HW \times f_s}$ and $\hat{Q} \in \mathbb{R}^{HW \times f_q}$, respectively. The core tensor $\mathcal{T}$ fuses $\hat{S}$ and $\hat{Q}$ together to generate activation confidence maps. M_o is a matrix which transfers the activation confidence maps to activated features ($A \in \mathbb{R}^{HW \times HW \times D_o}$), which are summarized and reshaped to obtain the intermediate activation maps, denoted as $\hat{A} \in \mathbb{R}^{D_o \times H \times W}$, Fig. 4.

The low-rank decomposition procedure aims to reduce the computational cost by reducing rank of the core tensor, $\mathcal{T}$. We decompose $\mathcal{T}$ into multiple slices (matrices) along dimension $\mathbb{R}^{f_o}$. Specifically, by setting rank of the slices (matrices) to $L \leq \min\{f_s, f_q\}$, we have the low-rank tensor $\mathcal{T}_L$. According to linear algebra, a L-rank matrix can be decomposed into L 1-rank matrices, and a 1-rank matrix can be represented by the outer product of a column vector and a row vector. As shown in Fig. 4, the k^{th} slice of $\mathcal{T}_L$ is decomposed by

$$\mathcal{T}_L[:,:,k] = \sum_{m=0}^{L} u_m^k \times (v_m^k)^\top, \tag{5}$$

where $u_m^k \in \mathbb{R}^{f_s \times 1}$ and $(v_m^k)^\top \in \mathbb{R}^{1 \times f_q}$ respectively denote a column vector and a row vector (a column-row vector pair).

Replacing A in Eq. 4 with A_L, we have

$$A_L = \mathcal{T}_L \times_1 \hat{S} \times_2 \hat{Q} \times_3 M_o. \tag{6}$$

which denotes activated features generated with the L-rank decomposition. L is experimentally determined to balance accuracy and computational cost. By defining

$$C_L = \mathcal{T}_L \times_1 \hat{S} \times_2 \hat{Q}, \tag{7}$$

we have

$$A_L = C_L \times_3 M_o. \tag{8}$$

Substituting Eq. 5 into Eq. 7, we have

$$C_L[:,:,k] = \sum_{m=0}^{L} \left(\hat{S} \times u_m^k\right) \times \left((v_m^k)^\top \times \hat{Q}^\top\right), \tag{9}$$

where $C_L \in \mathbb{R}^{HW \times HW \times f_o}$ denotes confidence maps produced by the bilinear activation procedure. $C_L[:,:,k]$ is the k-th slice of confidence map along dimension $\mathbb{R}^{f_o}$. Eq. 9 thereby defines dense correlation between $\hat{S}$ and $\hat{Q}$. Such correlation first compresses $\hat{S}$ and $\hat{Q}$ to vectors by u_m^k and $(v_m^k)^\top$ then fuses the compressed vectors to generate a confidence map by a product operation.

The procedure of bilinear activation with low-rank decomposition is detailed in Fig. 4. In the procedure, we first decompose the f_o slices to $f_o{*}L$ column-row vector pairs, and the column and row vectors are represented as two matrices $\in \mathbb{R}^{f_o \times L}$. Each column-row vector pair $(u_m^k, (v_m^k)^\top)$ multiplies with $(\hat{S}, \hat{Q})$, squeezing $(\hat{S}, \hat{Q})$ into semantic vectors. Given $f_o * L$ row-column pairs, we generate $f_o * L$ semantic vectors in total for the query and support images, respectively. We then multiply the support and query semantic vectors to generate a map matrix $\in \mathbb{R}^{f_o \times L}$. The elements of the map matrix are maps $\in \mathbb{R}^{HW \times HW}$. Following Eq. 9, we conduct summation and concatenation operations along the second and the first dimension of the map matrix. Then we obtain confidence maps ($C_L \in \mathbb{R}^{HW \times HW \times f_o}$). We follow Eq. 8 to multiply C_L with M_o, and get activated features ($A_L \in \mathbb{R}^{HW \times HW \times D_o}$), which is further summarized along the first dimension and reshaped to get the intermediate activation maps ($\hat{A}_L \in \mathbb{R}^{D_o \times H \times W}$). The core tensor T is learnable and randomly initialized. During training, it is updated at each iteration via back-propagation. During testing, it is fixed and used to fuse support and query features. In low-rank decomposition, $\mathcal{T}$ is decomposed into u_m^k and v_m^k. $M_s, M_q, M_o, u_m^k, v_m^k$ are also learnable matrices which are updated during training and fixed during testing.

C. Semantic Diffusion

Bilinear activation tends to activate object regions (*e.g.*, the head and tail of a cow) of large semantic similarity with the support image(s). Nevertheless, the object regions of small semantic similarity with the support image(s) could be unfortunately ignored. To pursue harmonic activation maps, a semantic diffusion procedure, which considers the local semantic consistency and intra-relevance of the query image, is proposed to refine the intermediate activation maps.

During semantic diffusion, pixels within the intermediate activation maps are updated according to its r^2 neighbors and diffusion weights. The process to construct diffusion weights P are illustrated in Fig. 5, where we use conv-blocks to encode

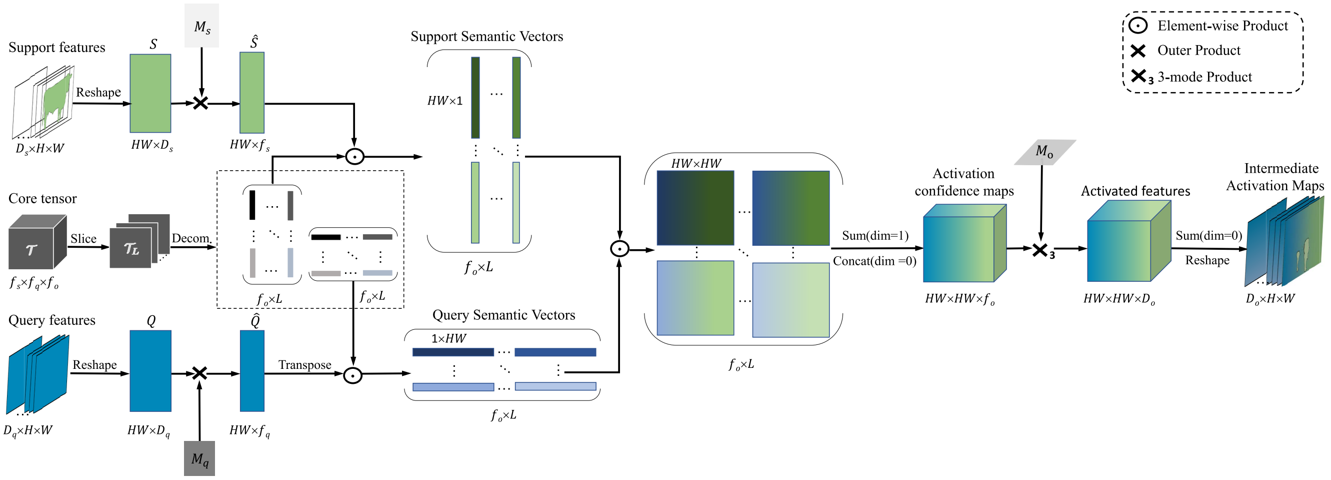


Fig. 4. Bilinear feature activation equipped with low-rank decomposition. We use M_s and M_q to project support and query features into a latent space, where the dimensionality of feature maps is significantly reduced. We then use the core tensor $\mathcal{T}$ to fuse the feature maps and generate activation maps. In the activation process, each slice of $\mathcal{T}$ is decomposed to L row-column vector pairs. The vector pairs are multiplied with the feature maps and calculate the activation confidences. (Best viewed in zoom).

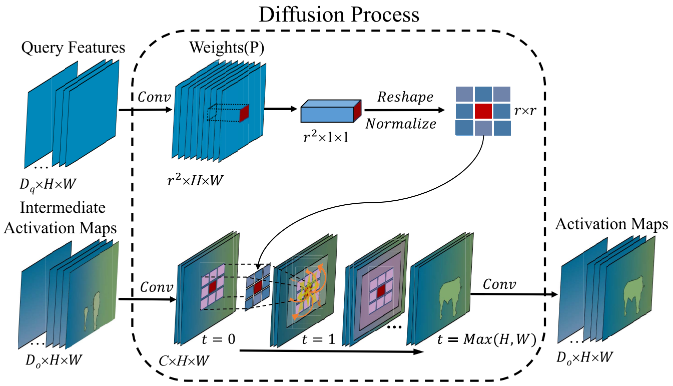


Fig. 5. The iterative diffusion process of semantic diffusion module. Diffusion weights are generated from query features by a convolution operation. The value of each pixel is updated with the values of its neighbors, and the intermediate activation maps are then iteratively updated based on diffusion weights.

Algorithm 1 Semantic Diffusion

```
Input: Intermediate activation maps $\hat{A}_L \in \mathbb{R}^{D_o \times H \times W}$, query
    features $Q_f \in \mathbb{R}^{D_q \times H \times W}$
Output: Activation maps $A_o \in \mathbb{R}^{D_o \times H \times W}$.
Define Img2Col($X$, $Y$): Expanding each sliding window
    of the image to a vector. $X, Y$ denote the height and width
    of the window, respectively.
Initialize diffusion weights $P = Conv(Q_f, \theta) \in \mathbb{R}^{9 \times H \times W}$, where $Conv$ is a stack of Conv-ReLU blocks
    and $\theta$ denotes learnable parameters;
Encode $\hat{A}_L$ and lower its dimension to $C \times H \times W$
Expand $P$ to $(3 \times 3) \times C \times (H \times W)$
Img2Col(3, 3) $\hat{A}_L$ to $(3 \times 3) \times C \times H \times W$
for $t = 1 : max(H, W)$ do
    $\hat{A}_L^t = P * \hat{A}_L^{t-1}$
end for
Reshape and encode $\hat{A}_L$ to $A_o \in \mathbb{R}^{D_o \times (H \times W)}$
return $A_o$
```

the query features and transfer its dimensionality to $\mathbb{R}^{r^2 \times H \times W}$. For each spacial location of the encoded maps, we extract a vector with the dimensionality $\mathbb{R}^{r^2 \times 1 \times 1}$, representing the diffusion weights of r^2 neighbors. Based on the weights, pixels within the same objects tend to have stronger connections, thus they are more likely to diffuse the activation confidence to each other. We then use conv-blocks to reduce the dimensionality of intermediate activation maps ($\hat{A}_L \in \mathbb{R}^{C \times H \times W}$), reducing the parameters and implementing efficient diffusion. Semantic diffusion iterates K steps, as

$$\hat{A}_L^t = D(\hat{A}_L^{t-1}, P), \tag{10}$$

$$P = Conv(Q, \theta) \in \mathbb{R}^{r^2 \times H \times W} \tag{11}$$

where $Conv$ denotes a stack of Conv-ReLU blocks and $\hat{A}_L^t$ denotes the activation maps generated by the t^{th} diffusion step, $0 < t \leq K$. D is the diffusion function. $P \in \mathbb{R}^{r^2 \times H \times W}$ denotes the diffusing weights, which are generated from query features Q_f, Eq. 11. θ denotes the learnable parameters of Conv-ReLU blocks.

Denote the i^{th} channel of the intermediate activation maps as $\hat{A}_{L,i}^t$. During each diffusion iteration, pixels of $\hat{A}_{L,i}^t$ are updated by the linear weighting of their r^2 neighbors with diffusing weights:

$$\hat{A}_{L,i;m,n}^t = \sum_{u,v \in N_{i;m,n}} Cons_{i;m,n} P_{i;m,n;u,v} \hat{A}_{L,i}^{t-1}, \tag{12}$$

where t denotes the t^{th} diffusion iteration. $N_{i;m,n}$ denotes the r^2 neighbors of coordinate (m,n), $Cons_{i;m,n}$ is a constraint which normalizes the diffusing weights. To guarantee that all pixels are updated, the maximum iteration number of the diffusion process is setted to $Max(H, W)$. In experiments, r is set to 3. The diffusion procedure is detailed in Algorithm 1.

D. Few-Shot Segmentation

As shown in Fig. 2, our few-shot segmentation network is constructed based on a metric learning framework with a support branch and a query branch. Following previous works [6]–[9], [11], [31], [36], [37], we use VGG-16 [38] or ResNet-50 [39] pre-trained on ImageNet as a backbone and use convolutional features from res-block/vgg-block 2 and 3 of the backbone ResNet-50/VGG-16 to generate feature maps. Following previous works [8], [10], the output stride of support and query feature maps is 8, which is 1/8 of the input image size. The support and query branches share a backbone network for feature extraction. We follow CANet [8] without iterative optimization and attention modules to construct the segmentation network.

During training or testing, the inputs of support branch are image-mask pairs. An element-wise product between support features and support ground-truth mask is used to filter out the background features. The preserved foreground features are used as semantic representation to guide the segmentation of the query image. After extracting foreground features of the query image, we use the HFA, *i.e.*, bilinear feature activation and semantic diffusion, to fuse the support and query features.

The segmentation network is trained in an end-to-end fashion driven by the Binary Cross Entropy (BCE) loss between the ground-truth and the segmentation mask. We randomly sample support-query pairs from the training set where the support and query images containing objects from the same categories. During inference, the fusion of the two modules makes full use of detailed semantic information in support and query images to achieve harmonic activation. The activation maps generated by the HFA approach together with query features are further processed by the query branch, segmenting the target object(s) after a few convolutional operations.

IV. Experiments

We first describe the experimental settings. We then report the performance of the proposed few-shot semantic segmentation approach and compare it with the state-of-the-art methods. We finally present ablation studies of the proposed approach.

A. Experimental Setting

1) Datasets: The experiments are conducted on MS COCO [40] and PASCAL VOC [41] datasets. For MS COCO, the experimental settings follow [10]. The dataset is divided into 4 splits, each of which contains 20 categories. For each split, 60 classes are used for training and the rest 20 classes for test. For each split, 1000 pairs of support and query images are randomly selected for performance evaluation. We combine the PASCAL VOC 2012 with SBD [42] and separate the combined dataset into 4 splits. Cross-validation is used by sampling five classes as test categories $C_{test} = 4i+1, \dots, 4i+5$, where i is the index of a split. The remaining 15 classes are used for training. During evaluation, 1000 pairs of support and query images are randomly selected to calculate the mean Intersection over Union (mIoU) and binary Intersection over Union (FBIoU) of all categories following previous works [9] [10] [8].

2) Training and Evaluation: For training, a batch size 8 is used. The weight decay is 1e-4 and momentum is 0.9. The segmentation model (network) is trained for 200000 steps with the poly descent training strategy and the stochastic gradient descent (SGD) optimizer. Data augmentation strategies including normalization, horizontal flipping, random rotation, random cropping and random resizing are adopted [8]. Following CANet [8], training images are resized to 321 × 321. Both single-scale and multi-scale evaluation [8] strategies are adopted for fair comparisons. For multi-scale evaluation, each image is augmented to multi-scale images by 0.7 and 1.3 times their original sizes and the predicted results of multi-scale images are averaged. To reduce randomness, we average mIoUs of multiple runs with different random seeds. Our approach is implemented with PyTorch 1.0 and run on Nvidia 2080Ti GPUs.

3) Evaluation Metric: The mIoU calculates the per-class foreground IoU and average the IoUs of all classes. The FB-IoU calculates the mean of foreground and background IoUs over all images regardless of the categories. We use both IoU and FB-IoU for evaluation. For category k, IoU is defined as $IoU_k = TP_k/(TP_k + FP_k + FN_k)$, where TP, FP and FN repsectively denote the numbers of true positives, false positives and false negatives. mIoU is the average of IoUs of all test categories and FB-IoU is the average of IoUs of all test categories and the background. mIoUs are averaged on four cross-validation splits.

B. Performance and Comparison

1) MS COCO: In Table I, we compare HFA with the state-of-the-art methods on MS COCO. HFA outperforms the state-of-the-art methods in 1-shot and achieves comparable results in 5-shot settings. Under 1-shot setting, it improves the baseline by 5.92%, respectively outperforming the PPNet and RPMMs methods by 3.81% and 2.42%. Under the 5-shot setting, it improves the baseline by 7.48%, respectively outperforms the PANet and FWB methods by 4.26% and 10.31%, which are significant margins for the challenging task.

2) PASCAL VOC: In Table II and Table III, we compare HFA with the state-of-the-art methods on Pascal VOC. HFA outperforms state-of-the-art methods under both 1-shot and 5-shot settings. Under 1-shot settings, with a VGG16 backbone, it respectively outperforms the FWB [10] and RPMMs [36] methods by 1.23% and 2.43%. Under the 1-shot settings, with a ResNet50 backbone, HFA outperforms the PPNet [37] method by 3.94% and achieves the new state-of-the-art. Under the 5-shot settings, HFA is comparable with the state-of-the-arts. Note that the PPNet and RPMMs used the additional k-shot fusion strategies while HFA uses a simple averaging strategy to fuse the five-shot results.

In Table IV, HFA is compared with the state-of-the-art approaches with respect to FB-IoU, which reflects how well full object extent is activated. Once gain, HFA outperforms the compared approaches under both 1-shot and 5-shot settings.

3) Segmentation Examples: In Fig. 6, we show segmentation examples by the baseline method and our HFA approach. These examples clearly demonstrate that HFA produces significantly better results while the baseline approach

TABLE I

PERFORMANCE OF 1-SHOT AND 5-SHOT SEMANTIC SEGMENTATION ON THE MS COCO DATASET. FWB USES THE RESNET101 BACKBONE WHILE OTHER APPROACHES USE THE RESNET50 BACKBONE.* DENOTES METHODS WITH MULTI-SCALE EVALUATION

Settings	Method	COCO-20^0	COCO-20^1	COCO-20^2	COCO-20^3	Mean
1-shot	PANet [9]	-	-	-	-	20.90
	FWB [10]	16.98	17.98	20.96	28.85	21.19
	PPNet [37]	**36.48**	26.53	25.99	19.65	27.16
	RPMMs* [36]	29.28	34.81	27.08	27.27	29.61
	Baseline	25.08	30.25	24.45	24.67	26.11
	HFA(ours)	27.53	**34.98**	**29.21**	**32.15**	**30.97**
	HFA*(ours)	28.65	**36.02**	**30.16**	**33.28**	**32.03**
5-shot	PANet [9]	-	-	-	-	29.70
	FWB [10]	19.13	21.46	23.93	30.08	23.65
	PPNet [37]	**48.88**	31.36	**36.02**	30.64	**36.73**
	RPMMs* [36]	33.82	41.96	32.99	33.33	35.52
	Baseline	25.95	32.38	26.11	26.98	27.86
	HFA(ours)	31.49	40.96	28.53	**34.87**	33.96
	HFA*(ours)	32.69	**42.12**	30.35	**36.19**	35.34

TABLE II

PERFORMANCE OF 1-WAY 1-SHOT SEMANTIC SEGMENTATION ON PASCAL-5^i. * DENOTES MULTI-SCALE EVALUATION

Backbone	Method	Pascal-5^0	Pascal-5^1	Pascal-5^2	Pascal-5^3	Mean	Params
VGG16	OSLSM [6]	33.60	55.30	40.90	33.50	40.80	272.6M
	co-FCN [31]	36.70	50.60	44.90	32.40	41.10	34.2M
	SG-One [7]	40.20	58.40	48.40	38.40	46.30	19.0M
	PANet [9]	42.30	58.00	51.10	41.20	48.10	14.7M
	CANet* [8]	-	-	-	-	54.30	-
	CRNet* [12]	-	-	-	-	55.20	-
	FWB [10]	47.04	59.64	52.51	48.27	51.90	43.0M
	RPMMs* [36]	47.14	65.82	50.57	48.54	53.02	-
	HFA(ours)	**48.05**	**65.95**	51.51	47.01	53.13	33.8M
	HFA*(ours)	**51.21**	**67.82**	**52.88**	**49.87**	**55.45**	33.8M
Resnet50	CANet* [8]	52.50	65.90	51.30	51.90	55.40	36.4M
	PGNet*[11]	**56.00**	66.90	50.60	50.40	56.00	32.5M
	CRNet*[12]	-	-	-	-	55.70	-
	PPNet[37]	48.58	60.58	**55.71**	46.47	52.84	31.5M
	RPMMs*[36]	55.15	66.91	52.61	50.68	56.34	-
	HFA(ours)	52.96	**68.97**	53.49	51.71	**56.78**	36.5M
	HFA*(ours)	53.89	**69.32**	54.52	**53.01**	**57.69**	36.5M

TABLE III

PERFORMANCE OF 1-WAY 5-SHOT SEMANTIC SEGMENTATION ON PASCAL-5^i. * DENOTES MULTI-SCALE EVALUATION

Backbone	Method	Pascal-5^0	Pascal-5^1	Pascal-5^2	Pascal-5^3	Mean
VGG16	OSLSM [6]	35.90	58.10	42.70	39.10	43.95
	SG-One [7]	41.90	58.60	48.60	39.40	47.10
	FWB [10]	50.87	62.86	56.48	50.09	55.08
	CRNet* [12]	-	-	-	-	**58.50**
	PANet [9]	51.80	64.60	**59.80**	46.05	55.70
	RPMMs* [36]	50.00	66.46	51.94	47.64	54.01
	HFA(ours)	**53.52**	**67.83**	52.60	49.98	55.98
	HFA*(ours)	**54.36**	**68.67**	53.54	**50.75**	**56.83**
Resnet50	CANet* [8]	55.50	67.80	51.90	53.20	57.10
	PGNet*[11]	**57.70**	68.70	52.90	54.60	58.50
	CRNet*[12]	-	-	-	-	58.80
	PPNet[37]	58.85	68.28	**66.77**	**57.98**	**62.97**
	RPMMs*[36]	56.28	67.34	54.52	51.00	57.30
	HFA(ours)	55.17	**69.99**	56.82	52.63	58.65
	HFA*(ours)	55.96	**70.35**	**57.28**	53.85	59.36

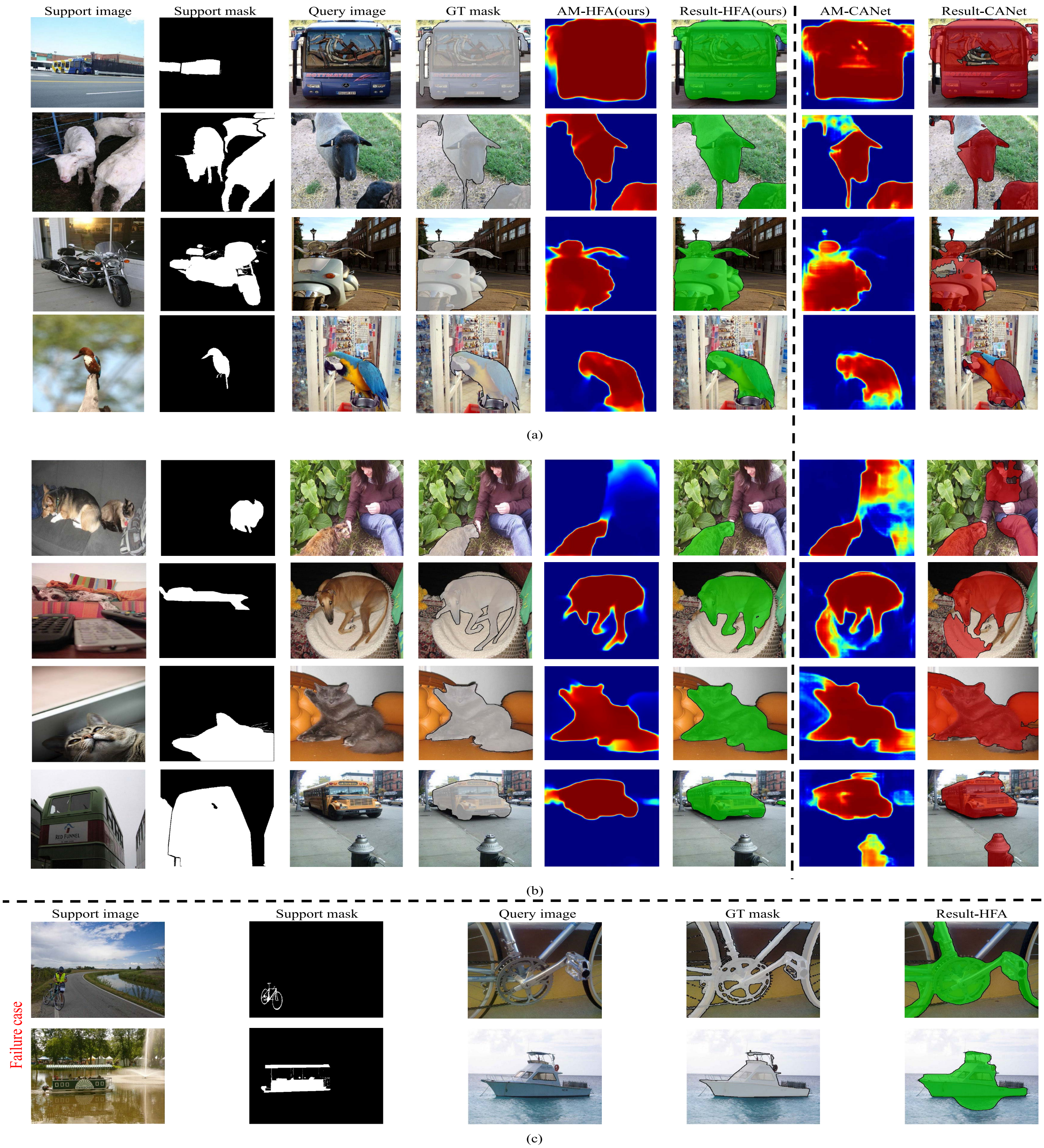
TABLE IV

COMPARISON OF FB-IOU PERFORMANCE OF 1-SHOT AND 5-SHOT SEGMENTATION ON THE PASCAL VOC 2012 DATASET

Method	1-shot	5-shot
Fine-tuning [43]	55.1	55.6
OSLSM [6]	61.3	61.5
co-FCN [31])	60.1	60.2
PL [44]	61.2	62.3
A-MCG [45]	61.2	62.2
SG-One [7]	63.9	65.9
PANet [9]	66.5	70.7
CANet [8]	66.2	69.6
PGNet [11]	69.9	70.5
CRNet [12]	66.8	71.5
HFA(ours)	**70.6**	**72.8**

tends to produce more missing segmentations, Fig. 6(a), and false labelings, Fig. 6(b). The good segmentation results of HFA are based on the sophisticated harmonic activation mechanism. Such a mechanism can fully leverage the semantic correlation between the support and query images and semantics within the query image to activate complete object extent.

4) Failure Cases: We present some failure cases in Fig. 6(c). The first failure example is a bicycle, which contains hollow regions. The strong activation on the gear of the bicycle is falsely propagated to the hollow regions by semantic diffusion. The second example is a boat, which is falsely segmented for the mirrored reflection regions connected with the true object.



(a)

(b)

(c)

Fig. 6. Segmentation examples. While the proposed HFA approach correctly segments the objects, the baseline CANet approach tends to produce missing segmentation (a) and false labeling (b). Failure examples (c) by HFA. The prefix ‘AM-’ refers to activation map.

C. Ablation Studies

1) Bilinear Feature Activation (BFA): In Table V, with BFA, we improve the segmentation performance by 3.94% (55.88% vs. 51.94%). We validate how BFA preserves the detailed semantic information from the perspective of activation accuracy. The sums of activation confidences within the target region (true positives) are defined as p_t, and those outside the target region (false positives) as p_f. Both p_t and p_f are scalars. We define the activation accuracy as $\frac{p_t - p_f}{p_t + p_f}$. Activation accuracy is calculated on feature maps with float values before the segmentation module, so that it can precisely reflect how well the semantic information is preserved and transferred. We sample 500 images per category and calculate the activation accuracy of activation maps generated

TABLE V

Ablation Study of the Proposed Approach on Pascal VOC. "Low-Rank Decom." Denotes Bilinear Feature Activation With Low-Rank Decomposition While "Bilinear" Denotes Bilinear Feature Activation Without Low-Rank Decomposition

Bilinear	Low-rank Decom.	diffusion	mIoU
			51.94
✓			55.88
	✓		55.62
		✓	54.82
✓		✓	56.81
	✓	✓	**56.78**

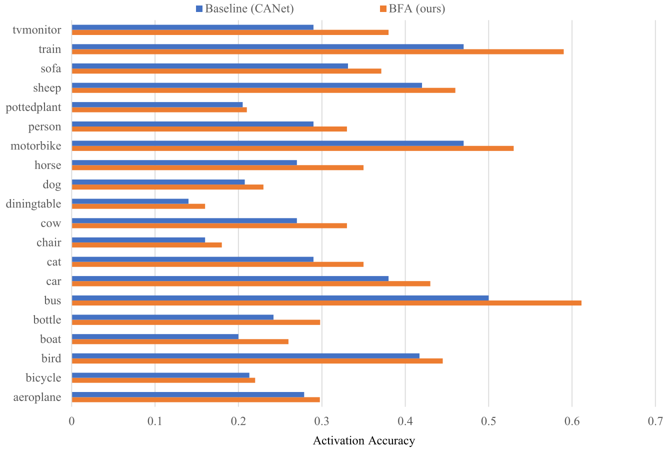


Fig. 7. Comparison of activation accuracy of bilinear feature activation (intermediate activation maps) and CANet.

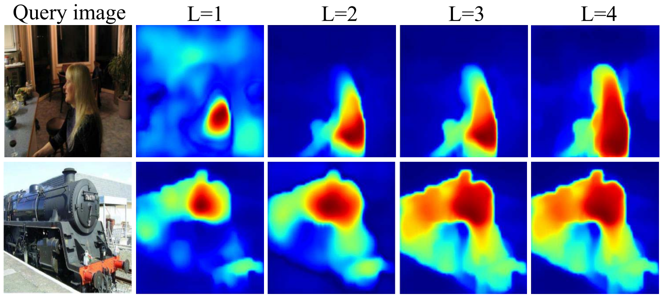


Fig. 8. Activation maps with different ranks (L).

by bilinear feature activation (intermediate activation maps) and CANet [8]. From Fig. 7 one can see that the activation accuracy of bilinear feature activation is significantly higher than that of CANet, validating it retains detailed semantic information of support features.

2) Low-Rank Decomposition: In Table V, when using the low-rank decomposition to approximate bilinear activation, the performance drop is negligible (0.26%), validating the effectiveness of Tucker decomposition. In Fig. 8, activation maps with different ranks (L) are presented. It can be seen that a larger rank contributes more complete activation. The impacts of rank on mIoU are shown in Fig. 9, according to which we set the rank $L = 3$.

3) Semantic Diffusion: This procedure also significantly improves the performance, as shown in Tab V. With semantic diffusion without bilinear activation, we improve the segmentation performance by 2.88% (54.82% vs. 51.94%). This shows that the intrinsic semantic consistency with the query image is important for few-shot segmentation, which is unfortunately ignored by existing works. In Fig. 10, activation maps under different iterations of diffusion are presented.

TABLE VI

Inference Time Per Example in Seconds. The Experiment Is Performed With a Single Nvidia-2080-Ti GPU on PASCAL-5i. $\Delta\theta$ Denotes the Increased Number(Ten Thousand) of Parameters Against Baseline(CANet)

Backbone	Baseline	HFA(Bilinear)	HFA(Low-rank)	HFA(Diffusion)	HFA
VGG	0.091	0.432	0.096	0.097	0.097
Res-50	0.157	0.621	0.162	0.162	0.163
$\Delta\theta$	0	98.6	6.3	1.5	7.8

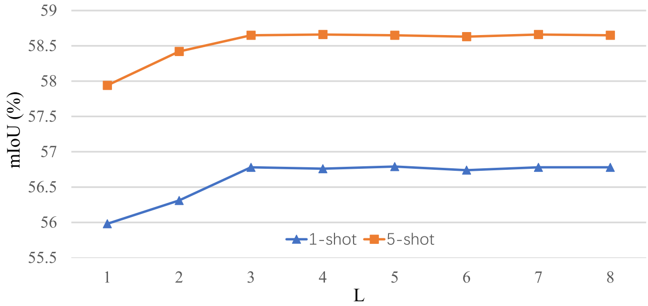


Fig. 9. mIoU with different ranks (L) on one-shot and five-shot settings. Experiments are conducted on the PASCAL VOC dataset.

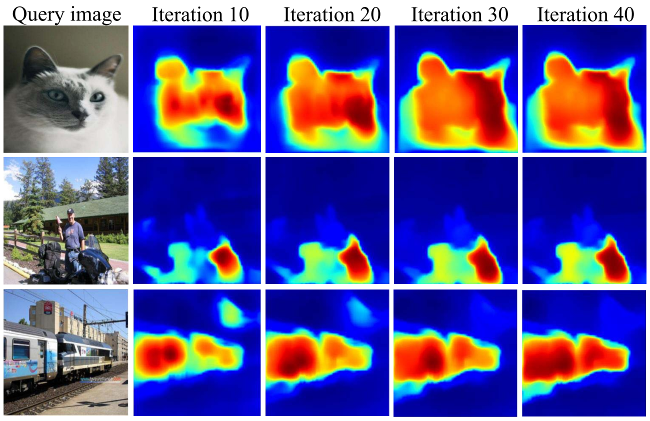


Fig. 10. Activation maps with different diffusion iterations. The maximum iteration number is 40.

4) HFA: With a combination of bilinear feature activation with semantic diffusion, our HFA approach improves the segmentation performance by 4.84% (56.78% vs. 51.94%), which validates that the two modules are complementary and can be fused to enforce harmonic feature activation.

5) Inference Time: In Table VI, the inference time of different modules are evaluated. With the VGG network, HFA (Low-rank) averagely uses 0.096 seconds to segment an image while HFA (Bilinear) uses 0.432 seconds. With the ResNet-50 network, HFA (Low-rank) averagely uses 0.162 seconds while HFA(Bilinear) uses 0.621 seconds. The 3.8 $\sim$ 4.5 times speed up shows that the proposed low-rank decomposition can significantly reduce the inference time with negligible performance cost. Meanwhile, the time cost of the semantic

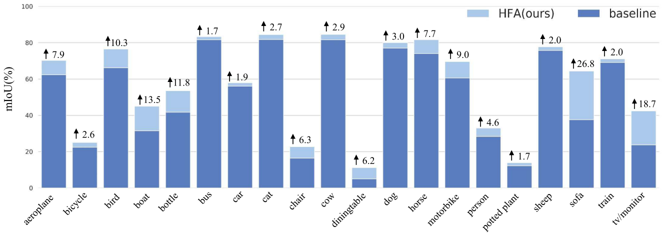


Fig. 11. Category-wised performance gains on the PASCAL VOC dataset. Our method(HFA) achieves significant improvements against the baseline(CANet).

TABLE VII

PERFORMANCE UNDER DIFFERENT LEARNING RATE SCHEDULES. "FIXED" DENOTES SETTING THE LEARNING RATE TO A FIXED VALUE, "REDUCE" DENOTES REDUCING THE LEARNING RATE BY ITERATIONS, AND "POLY" MEANS THE POLYNOMIAL LEARNING RATE

Schedule	Fixed	Reduce	Poly
mIoU	57.54	57.59	**57.73**

diffusion module is negligible. It uses only additional 0.001s (0.097 vs. 0.096).

6) Learning Rate Schedule: We compared commonly used learning rate schedules. The learning rate in PANet [9] is reduced by 0.1 every 10,000 iterations, and the learning rate in CANet [8] and FWB [10] is set to a fixed value. In Table VII, the polynomial learning rate schedule brings 0.14% performance gains against "Reduce" and 0.19% performance gains against "Fixed". When using the polynomial learning rate schedule, the testing results are more stable although it has negligible impact on the final performance.

7) Fusion Strategy: We compare different fusion strategies, including attention [6], cosine similarity [7] and concatenation [8] and find that bilinear feature activation achieves the best performance, Table VIII.

D. Statistic Analysis

1) Category-Wise Performance: In Fig. 11, we compare the category-wise segmentation performance on Pascal VOC. We sample 200 images from each category and calculate the performance gains. The categories of the largest performance gains are "sofa", "tv/monitor", "boat", and "bird". These categories can be largely affected by object views and poses. HFA has larger performance gains upon these categories, showing its potential to handle view and pose variations.

2) Object Size and mIoU: To further verify the effectiveness of the proposed approach, we analyze the relation between the sizes of target objects with mIoUs. We sample 2000 support-query pairs and calculate their sizes and mIoUs. In Fig. 12, HFA achieves averagely higher mIoU with respect to the distribution of object sizes. This attributes to the semantic diffusion procedure, which propagates semantics across object extent for complete segmentation.

TABLE VIII

PERFORMANCE UNDER DIFFERENT FUSION STRATEGIES

Method	Attention	Cosine	Concat.	BFA
mIoU	50.08	51.23	51.02	**53.13**

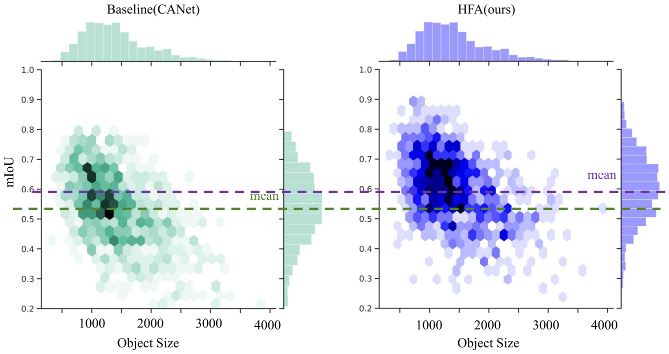


Fig. 12. Comparison of mIoUs over object size. The mIoU of HFA is averagely higher than that of the baseline approach.

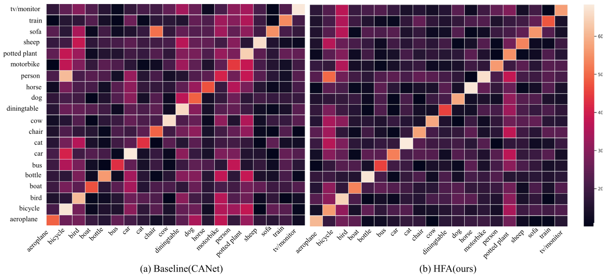


Fig. 13. Confusion matrix for the object categories on PASCAL VOC. HFA reduces the semantic confusion between categories and improves the model discriminability.

TABLE IX

MEAN-IOU PERFORMANCE OF 2-WAY 1-SHOT SEGMENTATION ON PASCAL VOC

Method	Pascal-5^0	Pascal-5^1	Pascal-5^2	Pascal-5^3	Mean
CANet [8]	47.03	60.66	46.32	46.93	50.24
PPN [37]	47.36	58.34	52.71	**48.18**	51.65
HFA (ours)	**49.02**	**64.31**	**49.13**	47.82	**52.57**

3) Model Discriminability: We sample and test 3500 images to draw the confusion matrix of object categories. As shown in Fig. 13, the x-coordinate indicates the ground-truth of samples while the y-coordinate indicates the predictions of the compared segmentation models. It can be seen that the baseline method tends to produce more false predictions for the categories including "boat", "motorbike", "person", and "dining table". When handling hollow and slender objects such as "bicycle" and "plotted plant", semantic diffusion might falsely propagate the activation confidence to hollow and/or slender regions (*e.g.*, gears, tyres, branches and leaves). HFA significantly reduces false predictions for most categories, demonstrating improved model discriminability.

E. Extended Experiments

1) 2-Way 1-Shot Segmentation: We have compared 2-way 1-shot segmentation performance with those of CANet [8] and PPNet [37], and PPNet is the state-of-the-art method under the 2-way 1-shot setting. We implement 2-way segmentation

TABLE X
PERFORMANCE OF MANET AND MANET+HFA

Method	AUC	J@60
MANet	0.749	0.761
MANet+HFA	**0.765**	**0.769**

by introducing a simple divide-and-conquer strategy. We first extract features from these 2 given support images, and use the support features to independently guide the segmentation of the query image to obtain segmentation results. We then concatenate the segmentation results and choose the class label with the highest confidence at each pixel location to implement 2-way segmentation. From Table IX, one can see that HFA achieves the best performance.

2) Video Object Segmentation: Our approach can be applied to few-shot video segmentation [46]–[49]. By using MANet [49] as the baseline, replacing the concatenation operation of MANet with the bilinear feature activation (BFA) module, and plugging the semantic diffusion module after BFA, we implemented video object segmentation. Experiment results in Table X show that HFA improves the performance of the baseline method by 1.6% (76.5% vs. 74.9%).

V. CONCLUSION

We proposed a novel few-shot segmentation approach, termed harmonic feature activation (HFA), and implemented precise support-to-query semantic transform by incorporating the features of both query and support images. HFA is formulated as a bilinear model, which takes charge of the pixel-wise dense correlation (bilinear feature activation) between query and support images. HFA incorporates a low-rank decomposition procedure, which greatly speeds up bilinear feature activation. A semantic diffusion procedure is fused with HFA, which further improves the global harmony and local consistency of the feature activation. HFA improved the performance of few-shot segmentation, in striking contrast with state-of-the-art approaches. The harmonic feature activation provides a fresh insight to the challenging few-shot learning problem.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention* (Lecture Notes in Computer Science), vol. 9351. Springer, 2015, pp. 234–241.

[2] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE CVPR*, Jul. 2017, pp. 6230–6239.

[3] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. ICLR*, 2015, pp. 6230–6239.

[4] M. Lin *et al.*, "HRank: Filter pruning using high-rank feature map," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1529–1538.

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, Feb. 2020.

[6] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–9.

[7] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "SG-one: Similarity guidance network for one-shot semantic segmentation," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3855–3865, Sep. 2020.

[8] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, "CANet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5217–5226.

[9] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "PANet: Few-shot image semantic segmentation with prototype alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 622–631.

[10] K. Nguyen and S. Todorovic, "Feature weighting and boosting for few-shot segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 622–631.

[11] C. Zhang, G. Lin, F. Liu, J. Guo, Q. Wu, and R. Yao, "Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9587–9595.

[12] W. Liu, C. Zhang, G. Lin, and F. Liu, "CRNet: Cross-reference networks for few-shot segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4165–4173.

[13] 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, Apr. 2018.

[14] L. Jing, Y. Chen, and Y. Tian, "Coarse-to-fine semantic segmentation from image-level labels," *IEEE Trans. Image Process.*, vol. 29, no. 7, pp. 225–236, 2020.

[15] M. Zand, S. Doraisamy, A. Abdul Halin, and M. R. Mustaffa, "Ontology-based semantic image segmentation using mixture models and multiple CRFs," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3233–3248, Jul. 2016.

[16] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. NeurIPS*, 2016, pp. 3630–3638.

[17] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.

[18] F. Hao, F. He, J. Cheng, L. Wang, J. Cao, and D. Tao, "Collect and select: Semantic alignment metric learning for few-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8460–8469.

[19] D. Das and C. S. G. Lee, "A two-stage approach to few-shot learning for image recognition," *IEEE Trans. Image Process.*, vol. 29, no. 12, pp. 3336–3350, 2020.

[20] S. Rahman, S. Khan, and F. Porikli, "A unified approach for conventional zero-shot, generalized zero-shot, and few-shot learning," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5652–5667, Nov. 2018.

[21] Y. Wang and M. Hebert, "Learning to learn: Model regression networks for easy small sample learning," in *Proc. ECCV*, 2016, pp. 616–634.

[22] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *Proc. ICLR*, 2017.

[23] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. ICML*, 2017, pp. 1126–1135.

[24] M. A. Jamal and G.-J. Qi, "Task agnostic meta-learning for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 111719–111727.

[25] B. Hariharan and R. Girshick, "Low-shot visual recognition by shrinking and hallucinating features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3037–3046.

[26] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan, "Low-shot learning from imaginary data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7278–7286.

[27] W.-Y. Chen, Y.-C. Liu, Z. Kira, and Y. C. Wang, "A closer look at few-shot classification," in *Proc. ICLR*, 2019.

[28] X. Liu, F. Zhou, J. Liu, and L. Jiang, "Meta-learning based prototype-relation network for few-shot classification," *Neurocomputing*, vol. 383, pp. 224–234, Mar. 2020.

[29] W.-H. Chu, Y.-J. Li, J.-C. Chang, and Y.-C.-F. Wang, "Spot and learn: A maximum-entropy patch sampler for few-shot image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6251–6260.

[30] N. Dvornik, C. Schmid, and J. Mairal, "Selecting relevant features from a universal representation for few-shot classification," *CoRR*, vol. abs/2003.09338, pp. 1–9, Mar. 2020.

[31] K. Rakelly, E. Shelhamer, T. Darrell, A. A. Efros, and S. Levine, "Conditional networks for few-shot semantic segmentation," in *Proc. ICLR Workshop*, 2018.

[32] D. López-Sánchez, A. González Arrieta, and J. M. Corchado, "Compact bilinear pooling via kernelized random projection for fine-grained image categorization on low computational power devices," *Neurocomputing*, vol. 398, pp. 411–421, Jul. 2020.

[33] T.-Y. Lin and S. Maji, "Improved bilinear pooling with CNNs," in *Proc. Brit. Mach. Vis. Conf.*, 2017.

[34] T. Doghri, L. Szczecinski, J. Benesty, and A. Mitiche, "Bilinear models for machine learning," *CoRR*, vol. abs/1912.03354, pp. 1–8, Sep. 2019.

[35] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, Aug. 2009.

[36] B. Yang, C. Liu, B. Li, J. Jiao, and Q. Ye, "Prototype mixture models for few-shot semantic segmentation," in *Proc. ECCV*, 2020, pp. 763–778.

[37] Y. Liu, X. Zhang, S. Zhang, and X. He, "Part-aware prototype network for few-shot semantic segmentation," in *Proc. ECCV*, 2020, pp. 142–158.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, Y. Bengio and Y. LeCun, Eds., Sep. 2015.

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

[40] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[41] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

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

[43] K. Rakelly, E. Shelhamer, T. Darrell, A. A. Efros, and S. Levine, "Conditional networks for few-shot semantic segmentation," in *Proc. ICLR*, 2018.

[44] N. Dong and E. P. Xing, "Few-shot semantic segmentation with prototype learning," in *Proc. BMVC*, 2018, p. 79.

[45] T. Hu, P. Yang, C. Zhang, G. Yu, Y. Mu, and C. G. M. Snoek, "Attention-based multi-context guiding for few-shot semantic segmentation," in *Proc. AAAI*, 2019, pp. 8441–8448.

[46] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5320–5329.

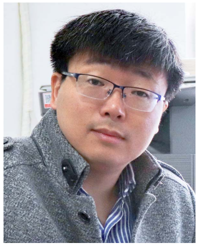
[47] P. Voigtlaender, Y. Chai, F. Schroff, H. Adam, B. Leibe, and L. Chen, "FEELVOS: Fast end-to-end embedding learning for video object segmentation," in *Proc. IEEE CVPR*, Jun. 2019, pp. 9481–9490.

[48] C. Liang, Z. Yang, J. Miao, Y. Wei, Y. Yang, "Memory aggregation networks for efficient interactive video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10366–10375.

[49] J. Miao, Y. Wei, Y. Yang, "Memory aggregation networks for efficient interactive video object segmentation," in *Proc. IEEE CVPR*, Jun. 2020, pp. 10366–10375.

Binghao Liu received the B.S. degree from Wuhan University, Wuhan, China, in 2018. He is currently pursuing the master's degree with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing, China. His research interests include computer vision and machine learning, specifically for representation learning and few-shot learning.

Jianbin Jiao (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in mechanical and electronic engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 1989, 1992, and 1995, respectively. From 1997 to 2005, he was an Associate Professor with HIT. Since 2006, he has been a Professor with the University of Chinese Academy of Sciences, Beijing, China. His research interests include image processing and pattern recognition.



Qixiang Ye (Senior Member, IEEE) received the B.S. and M.S. degrees from the Harbin Institute of Technology, China, in 1999 and 2001, respectively, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2006. He has been a Professor with the University of Chinese Academy of Sciences since 2009. He was a Visiting Assistant Professor with the Institute of Advanced Computer Studies (UMIACS), University of Maryland, College Park, MD, USA, until 2013. He has published more than 100 articles in conferences and journals, including CVPR, ICCV, ECCV, NeurIPS, IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), and IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (T-PAMI). His research interests include visual object detection and machine learning. He received the Sony Outstanding Paper Award.