# Few-Shot Semantic Segmentation with Democratic Attention Networks

Haochen Wang[1,6*], Xudong Zhang[1*], Yutao Hu[1], Yandan Yang[1], Xianbin Cao[1,2,3†], and Xiantong Zhen[4,5]

[1] Beihang University, Beijing, China
[2] Key Laboratory of Advanced Technology of Near Space Information System, Ministry of Industry and Information Technology of China
[3] Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, China
[4] AIM Lab, University of Amsterdam, The Netherlands
[5] Inception Institute of Artificial Intelligence, Abu Dhabi, UAE
[6] YouKu Cognitive and Intelligent Lab, Alibaba Group
{haochenwang,xdzhang,huyutao,yangyandan}@buaa.edu.cn, xbcao@buaa.edu.cn, zhenxt@gmail.com

**Abstract.** Few-shot segmentation has recently generated great popularity, addressing the challenging yet important problem of segmenting objects from unseen categories with scarce annotated support images. The crux of few-shot segmentation is to extract object information from the support image and then propagate it to guide the segmentation of query images. In this paper, we propose the Democratic Attention Network (DAN) for few-shot semantic segmentation. We introduce the democratized graph attention mechanism, which can activate more pixels on the object to establish a robust correspondence between support and query images. Thus, the network is able to propagate more guiding information of foreground objects from support to query images, enhancing its robustness and generalizability to new objects. Furthermore, we propose multi-scale guidance by designing a refinement fusion unit to fuse features from intermediate layers for the segmentation of the query image. This offers an efficient way of leveraging multi-level semantic information to achieve more accurate segmentation. Extensive experiments on three benchmarks demonstrate that the proposed DAN achieves the new state-of-the-art performance, surpassing the previous methods by large margins. The thorough ablation studies further reveal its great effectiveness for few-shot semantic segmentation.

**Keywords:** Few-Shot Segmentation, Graph Attention, Democratic Attention Network, Multi-Scale Guidance

## 1 Introduction

Deep convolutional neural networks driven by large-scale labeled datasets [3, 17] have shown great success in many visual recognition tasks, such as image classi-

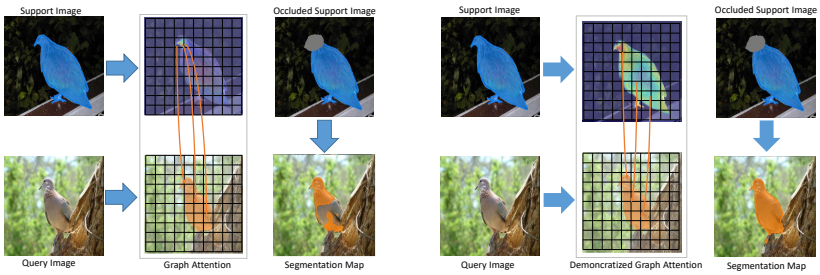
*These authors contribute equally.
†Corresponding author.

fication [14, 30, 9, 12] and semantic segmentation [19, 15, 8]. As for conventional semantic segmentation, training a deep neural network requires pixel-level annotation, which is costly and time consuming. In addition, once the model is learned, it is difficult to predict new classes absent in the training set. In contrast to machine learning models, humans are good at recognizing a new object even with a little guidance. Inspired by this, few-shot semantic segmentation has recently received growing interest in the computer vision community [4, 28, 39]. Few-shot semantic segmentation targets at learning transferable knowledge by segmenting objects of seen categories to generalize to new categories of objects, where only a few annotated support images are available.

Most of the current few-shot segmentation methods [28, 41, 25, 40, 4] are based on prototype learning, employing a two-branch encoder-decoder architecture, i.e., a support branch and a query branch. The support branch is deployed to extract a class prototype from the support images and the query branch takes the prototype as guidance for segmenting the query image. To obtain this guidance, [41, 36, 22] adopted global average pooling by squeezing the support feature of the support image into a vector, and based the segmentation on a specific metric, e.g., cosine similarity, between the global vector and the feature map of the query image. However, the mask average pooling operation inevitably drops the spatial information of the support images, leading to a noisy output. To solve this problem, [39] established the pixel-to-pixel connection between support and query images by leveraging the graph attention mechanism [37, 34]. Nonetheless, usually only a small region of the foreground object is activated in the support image due to the biased competition among the pixels, resulting in the connection between the support and query images being dominated by a small portion of pixels, largely limiting the information to be propagated. As illustrated in Fig. 1 (a), only pixels on the head region of the bird are activated. This would lead to overfitting and reduced robustness when foreground objects are partially occluded.

Furthermore, previous methods always merge the prototype and the feature map from the query branch at the highest semantic level. However, both the low-level features and the high-level features are essential for the segmentation procedure. For instance, to precisely segment a car in a query image, we require the guidance of high-level prototypes such as wheel characteristic, as well as the low-level prototypical information, such as the surface texture. A single prototypical vector would not be able to fully capture these different levels of semantic information. Therefore, it becomes essential to explicitly explore and retain different semantic levels of the prototypical information in the support image for the segmentation of query images.

In this paper, to deal with the two aforementioned problems, we propose the Democratic Attention Network (DAN) for few-shot semantic segmentation. We introduce a democratized graph attention (DGA) mechanism to establish the pixel-to-pixel correspondence between support and query images. The idea is to suppress the connections with high weights, while enhancing those with lower weights during training. This endows the network with the ability to establish




(a) Graph Attention (b) Democratized Graph Attention

**Fig. 1.** Graph attention (a) vs. democratized graph attention (b). The regular graph attention (GA) mechanism makes the network attentive to a dominate regains (e.g., the head of the bird in this example), which compromises the model robustness to noise, e.g., occlusion. Our DGA mechanism enables more pixels to be activated, establishing robust connections between support and query images.

robust connections by attending to larger regions of the object instead of only small specific regions. As a result, pixels on the foreground object tend to be democratized to participate in the connections. In this way, the network can propagate more guiding information from support and query images for segmentation. Once trained, the model acquires the ability to leverage more pixels from the foreground object in the support images, which enhances its robustness to occlusion and generalizability to new objects. As shown in Fig. 1 (b), our proposed DGA activates the larger regions, not only the bird's head region, resulting in a more robust connection, and thus making it insensitive to the partial occlusion of foreground objects in contrast to the regular graph attention [39].

Moreover, we propose a multi-scale guidance for the segmentation of query images by constructing a hierarchical graph attention to explore multi-level semantic information. Specifically, we apply the democratized graph attention mechanism to multiple intermediate feature layers in the encoder to establish the multi-level connection between query and support images. Coupled with the hierarchical graph, we design a refinement fusion unit in the decoder to fuse the multi-level attentive information with the corresponding feature layers of the query image in the decoder for improved segmentation.

We conduct extensive experiments on three benchmarks, i.e., the PASCAL-$5^i$ [28], COCO-$20^i$ [11] and FSS-1000 [38] datasets, for performance evaluation. To the best of our knowledge, this is the first work that provides performance evaluation on three datasets for few-shot segmentation. The proposed DAN largely surpasses previous state-of-the-art methods. Moreover, we perform thorough ablation studies to demonstrate the benefits of the proposed democratized graph attention and the multi-scale guidance, offering more insight into its great effectiveness for few-shot semantic segmentation.

The major contributions of this work are summarized as follows:

- We propose a new, democratic attention network (DAN) for few-shot semantic segmentation. The introduced democratized graph attention mechanism can activate more pixels on the foreground object, achieving stable correspondence between a support and query image. This enables more guiding information to be propagated from support to query images, enhancing robustness and generalizability to new objects.
- We introduce multi-scale guidance by designing a refinement fusion unit to efficiently fuse multi-level semantic information of the support image to improve the segmentation of query images.
- We achieve new state-of-the-art performance on three benchmarks, largely advancing the current performance of few-shot semantic segmentation.

## 2 Related Work

### 2.1 Semantic Segmentation

Semantic segmentation is one of the most popular tasks in computer vision, which aims to classify each pixel of an image into predefined categories. Previous works [16, 27, 2, 1] are typically based on a fully convolutional network (FCN) [19]. FCN replaces the fully connected layer with the convolution layer, which facilitates dense prediction and thus improves segmentation performance. Ronneberger et al. [27] proposed an encoder-decoder structure to produce the segmentation map with high resolution, in which the encoder network was designed to capture abstract feature representations and the decoder network is used to map the low-resolution encoder feature maps to full input resolution feature maps for pixel-wise classification. However, the traditional fully supervised segmentation methods require large amounts of pixel-level annotated images for training, which are expensive and time consuming to obtain. Additionally, once the model is trained, it cannot be generalized to unseen categories of objects.

### 2.2 Few-shot Learning

Few-shot learning, which learns new concepts from a few annotated examples, has recently generated great popularity. The existing few-show learning works focus on image classification tasks, aiming to predict image class labels given only a few training examples in each category. Few-shot learning has been investigated under the meta-learning framework by exploring shared metric/similarity space [31, 35, 33, 42], or optimization algorithms [26, 6, 23]. Few shot learning has also been explored in video analysis [20, 21]. Recently, few-shot classification has been extended to few-shot semantic segmentation [4], a more challenging task that aims to segment objects of unseen categories with limited annotated images.

### 2.3 Few-shot Semantic Segmentation

Existing deep models for few-shot semantic segmentation are mainly built upon a two-branch structure [28, 25, 41, 40, 11] which includes a support branch and

a query branch. The support branch aims to extract information from the support images, which could provide guidance to the query branch for predicting segmentation maps. Most previous prototypical methods [28, 40, 41, 36] extract a global vector from the support images to learn a class representation with limited support images. For instance, Zhang et al. [41] proposed a global average pooling operation to obtain the global vector and utilized a cosine similarity to build the relationship between support and query images. The extracted global vector can be upsampled and concatenated with the feature map of the query image to produce the segmentation map [40]. However, squeezing the support images to a global descriptor does not retain the spatial structure of the support images, which is crucial for the segmentation task. To solve this problem, Zhang et al. [39] proposed the construction of graphs to establish element-to-element connections to propagate information from the support image to query images. However, due to the limited annotation, the connections tend to be dominated by a small portion of pixels in the support images, caused by the biased competition among pixels. This compromises the robustness of the final prediction and reduces the generalization to new object classes. Moreover, they built a pyramid structure by deploying adaptive average pooling to the query feature map at the same semantic level, while the pyramid query feature maps only attend to the last single feature map of the support image, which fails to utilize multi-level semantic features.

In this work, we introduce a democratized graph attention mechanism to establish robust pixel-to-pixel connections between support and query images without falling into the centralized problem. This endows the network with the ability to extract robust categorical information from support images by constraining the connections with high weights and enhancing the remaining part. Further, we propose the multi-scale guidance by constructing hierarchical graph attentions at intermediate layers of the encoder, which offers multiple levels of semantic information to better guide the segmentation of query images.

## 3 Democratic Attention Network

### 3.1 Problem Definition

For a $k$-shot semantic segmentation task, the goal is to train a model to perform segmentation on new classes with scarce annotated images. Suppose we are provided with two image sets $D_{\text{train}}$ and $D_{\text{test}}$, where the $D_{\text{train}}$ is used for training the model and the $D_{\text{test}}$ is for evaluation. Note that, in contrast to conventional semantic segmentation, there is no overlap between the categories in $D_{\text{train}}$ and those in $D_{\text{test}}$. Both the training set $D_{\text{train}}$ and testing set $D_{\text{test}}$ are composed of several episodes. Each episode contains a support set $\mathcal{S}$ and a query set $\mathcal{Q}$, where $\mathcal{S} = \{\mathbf{x}_i^s, \mathbf{m}_i^s\}_{i=1}^k$ contains $k$ images $\mathbf{x}^s$ and corresponding binary masks $\mathbf{m}^s$ for a certain category $c$, and $\mathcal{Q} = \{\mathbf{x}^q, \mathbf{m}^q\}$ contains query image $\mathbf{x}^q$ to be segmented and the associated ground truth mask $\mathbf{m}^q$.

The network is trained by episodically sampling support and query pairs from $D_{\text{train}}$ to learn the mapping from $\{S, \mathbf{x}^q\}$ to the object mask $\mathbf{m}^q$. Note

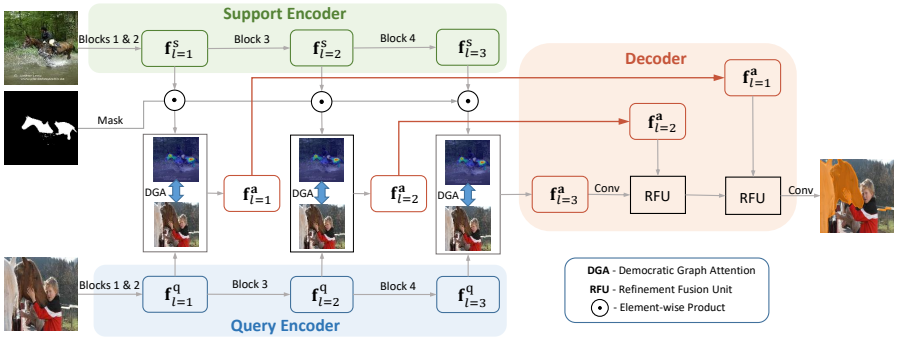



**Fig. 2.** Architecture of Democratic Attention Network (DAN) (illustrated in one-shot setting). Let $\{\mathbf{f}_l^s\}_{l=1}^L$ and $\{\mathbf{f}_l^q\}_{l=1}^L$ be the feature maps extracted in $L(=3)$ different layers of encoders for support and query images, respectively. The proposed democratized graph attention (DGA) mechanism is applied to those feature maps to establish the correspondence between support and query images. The obtained attentive feature maps $\{\mathbf{f}_l^a\}_{l=1}^L$ of different semantic levels are fed into the designed refinement fusion unit (RFU) to achieve multi-scale guidance for the segmentation of query images.

that once the network is learned, the mapping is fixed and requires no iterative optimization when tested on a test dataset $D_{\text{test}}$. The training procedure is set to be aligned with that of evaluation. Given a query image $\mathbf{x}^q$ and a support image-mask pair $(\mathbf{x}^s, \mathbf{m}^s)$ as input, our goal is to produce the segmentation map $\hat{\mathbf{m}}^q$ for query image.

### 3.2 Architecture Overview

As shown in Fig. 2, the proposed democratic attention network (DAN) is built upon a two-branch architecture, which is now widely used in existing few-shot segmentation networks [41, 40, 25]. We introduce two major innovative architectures into the two-branch network: (1) A democratized graph attention (DGA) mechanism is proposed to establish pixel-to-pixel connections between query and support images based on the graph attention. DGA constrains the connections with high weights while enhancing those with low weights during training, which endows the network with the ability to establish robust connections between support and query images. (2) We introduce a multi-scale guidance architecture by designing a refinement fusion unit, which enables multi-level semantic information into the guidance of segmenting query images.

To be more specific, we first deploy a weights-shared convolutional neural network as the feature extractor to acquire a sequence of deep features maps $(\{\mathbf{f}_l^q, \mathbf{f}_l^s\}_{l=1}^L)$ representing different semantic levels of information for the query image and the support image respectively, as shown in Fig. 2. Then, each pair of deep features $(\mathbf{f}_l^q, \mathbf{f}_l^s)$ is fed into the proposed democratized graph attention

block to establish the connection between the support and query images at each of the individual semantic levels. This results in hierarchical attentive maps $\{\mathbf{f}_l^a\}_{l=1}^L$ in multiple semantic levels, which are finally fused with features in the corresponding decoder layers by the designed refinement fusion unit to produce segmentation maps.

### 3.3 Democratized Graph Attention

One of the keys to few-shot semantic segmentation is to extract and propagate object information from the support image to the query image. Instead of using a prototype vector that loses the essential structure information [41, 40, 29], we establish pixel-to-pixel dense connections between the query image and the support image based on the graph attention [34]. However, the connection provided by the regular graph attention mechanism tends to be dominated by the small most discriminative region, which is not robust and lacks generalizability to new classes of objects. We introduce the democratized graph attention (DGA) mechanism to enhance the robustness of the connection, which is easy to implement and effectively improve the performance. Moreover, in contrast to the graph attention in [39], we propose to construct a hierarchical graph based on multiple levels of semantic features, which enables more guiding information to be propagated from the support to query image for more accurate segmentation.

To be more specific, we establish the pixel-to-pixel correspondence between the query image and the support image at multiple intermediate layers of the encoder network by our democratized graph attention mechanism. In each intermediate layer, the DGA takes the feature map $\mathbf{f}^s$ from the support image and the feature map $\mathbf{f}^q$ from the query image as input. As shown in Fig. 3, two convolutional layers are applied to embed the feature maps $\mathbf{f}$ into key maps $\mathbf{k}$ and value maps $\mathbf{v}$ separately, where the $\mathbf{k}$ are used to measure the correspondence between query and support images, and the $\mathbf{v}$ restore the extracted detailed information of the feature maps. Once we obtain $\mathbf{k}^s$, $\mathbf{v}^s$ for the support image and $\mathbf{k}^q$, $\mathbf{v}^q$ for the query image, the pixel-wise connection is constructed by estimating the graph affinity between $\mathbf{k}^q$ and $\mathbf{k}^s$.

The graph affinity $A$ is computed by measuring the similarities between all pixels of the query key map $\mathbf{k}^q$ and the support key map $\mathbf{k}^s$ with a pairwise function $g(\cdot,\cdot)$. The connection weight between the pixel $i$ on the query image and the pixel $j$ on the support image can be denoted as:

$$A_{i,j} = g\left(\mathbf{k}_i^q, \mathbf{k}_j^s\right) = \left(\mathbf{k}_i^q\right)^\top \times \mathbf{k}_j^s. \tag{1}$$

By applying $g(\cdot,\cdot)$ to each of the pixel pairs between the query and support image, we obtain the connection graph $A \in \mathbb{R}^{HW \times HW}$. We take the average of the graph affinity $A$ across the first dimension to produce the attention map for the support image $A^s = \sum_{i=1}^{HW} A_i \in \mathbb{R}^{H \times W}$, in which the value measures the average connection weights between each pixel in the support image and all pixels in the query image. The attention map $A^s$ generally reflects the importance of pixels in the support image in guiding the segmentation of objects on the query

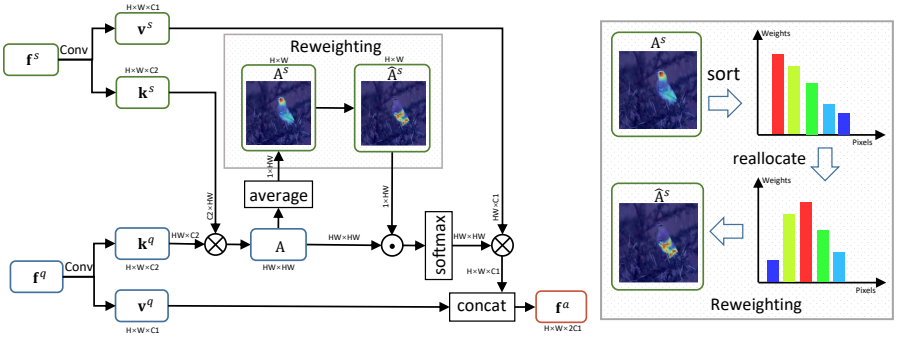



**Fig. 3.** Decentralized Graph Attention (DGA). The DGA weighs down the highly activated regions during training, which enables the network to leverage more informative pixels on the object for improved segmentation during testing.

image. However, in the regular graph attention, activated pixels tend to fall in a small discriminative region, which dominates the connections. As shown in Fig. 3, the most of the connection weights gather on the head of the bird, caused by the biased competition among the support pixels, which limits the information to be propagated from the support to query image.

To guarantee the generalization to new objects of great variations, it is desired to leverage as many pixels as possible on the annotated object of the support image. To this end, we design a new, democratized graph attention mechanism to establish better attention maps. Specifically, we propose sorting and reallocating the affinities of each pixel in $A^s$ during training. We sort the pixels of $A^s$ in descending order and acquire the sorting index $\mathbf{e}_j$ for each pixel $j$, and then we reallocate the connection weights to pixel $j$ according to the corresponding index by the function $\phi(A^s_j, \mathbf{e})$.

$$\hat{A}^s_j = \phi(A^s_j, \mathbf{e}) = A^s_j \times \frac{\mathbf{e}_j}{\frac{1}{HW}\sum_{j=1}^{HW} \mathbf{e}_j}, \tag{2}$$

where the $H, W$ indicate the height and width of $A^s$. After this operation, the pixels with high connection weights are suppressed so that the other part of pixels are enhanced as shown in Fig. 3, where the body of the bird are activated. The reconstructed attention map $\hat{A}^s$ is used to compute the weighted graph affinity:

$$\hat{A}_{i,j} = g\left(\mathbf{k}^q_i, \mathbf{k}^s_j \cdot \hat{A}^s_j\right) = (\mathbf{k}^q_i)^{\mathrm{T}} \times (\mathbf{k}^s_j \cdot \hat{A}^s_j). \tag{3}$$

The activated pixels will expand to less discriminative regions of the foreground object and more pixels will contribute to the information propagation, which gains more robustness and generalization ability. In a similar spirit to dropout [32], the DGA is only applied to a portion of samples during training. As a result,

the model learns to acquire the ability to use more pixels of the foreground object in the support images during the segmentation of new objects.

In addition, to make connection weights comparable across different pixels, we normalize them with the softmax function and generate the normalized graph affinity $\hat{A}^{'}$. Then the support value map $\mathbf{v}^s$ is fused by a weighted summation with the normalized graph affinity $\hat{A}^{'}$ and then concatenated with the query value map $\mathbf{v}^q$ to generate the attentive feature map $\mathbf{f}^a$:

$$\mathbf{f}_i^a = \{\mathbf{v}_i^q || \sum_j \hat{A}_{i,j}^{'} \cdot \mathbf{v}_j^s\}, \quad \hat{A}_{i,j}^{'} = \frac{\exp(\hat{A}_{i,j})}{\sum_j \exp(\hat{A}_{i,j})}, \tag{4}$$

where $\{\cdot||\cdot\}$ denotes the concatenation operation. Note that once we obtain the output attentive feature map of position $i$, the other positions could also be computed by applying the same operations as above.

### 3.4 Multi-Scale Guidance

The other key step is to fully utilize the information extracted from the support image to guide the segmentation of query images. We apply the DGA blocks at the different semantic levels to generate multiple attentive feature maps $\{\mathbf{f}_l^a\}_{l=1}^L$ that contain different levels of semantic information of the foreground object. To efficiently leverage these attentive feature maps, we design a refinement fusion unit to fuse the multi-level information in $\{\mathbf{f}_l^a\}_{l=1}^L$ with the corresponding decoder layers in a sequential manner.

As shown in Fig. 2, the refinement fusion unit in the decoder network upsamples its input representation map using a bi-linear upsampling operation and the obtained representation map is concatenated with the corresponding attentive feature map processed by a residual block. The concatenated feature map is then processed by a convolutional block to produce a dense representation map. The deepest output of the final refinement fusion unit is fed into a convolutional layer and a softmax operation is applied to distinguish each pixel independently. The output of the softmax layer is a two-channel map of probabilities that indicate foreground and background, respectively. We obtain the predicted segmentation map of the query image by taking the label of the corresponding class with a maximum probability at each pixel.

## 4 Experiments

**Implementation details.** The backbone architecture employed in our segmentation network is a resnet101 [10] that is pre-trained on ImageNet [3]. The loss function is the mean of cross-entropy loss over all pixel locations in the output segmentation map. The model is trained end-to-end by the Adam optimizer [13] using a batch size of 4 for 50000 iterations on an GeForce RTX 2080 Ti GPU. The learning rate is fixed to 1e-5 during training. We conduct multi-scale inference, where the scale rates are set to $\{0.5, 1, 1.5\}$ for both query and support images. For the $k$-shot setting, we concatenate the key maps and value maps produced by individual shots and acquire unified key and value maps.

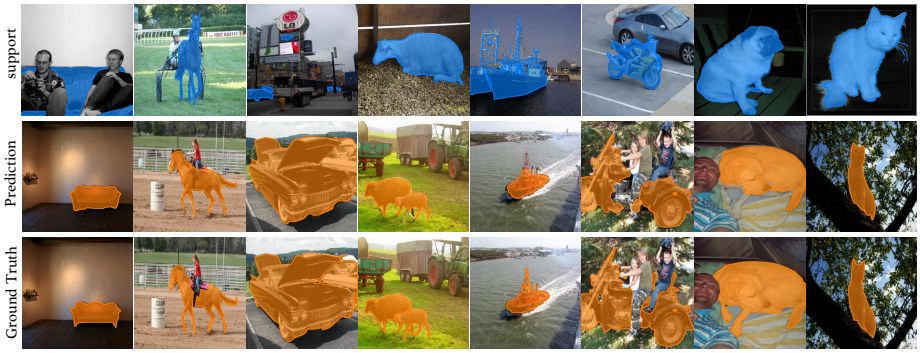



**Fig. 4.** Visualization of segmentation results on Pascal-$5^i$. Our DAN can make accurate segmentation even when query objects exhibit great variations from support ones.

**Evaluation metrics.** We take the commonly-used evaluation metrics, Mean-IoU and FB-IoU, to benchmark with previous methods. Mean-IoU in [28] is defined as the average per-class foreground Intersection-over-Union (IoU) over all classes, i.e., IoU $= \frac{tp}{tp+fp+fn}$, where $tp$ is the number of true positives, $fp$ is the number of false positives and $fn$ is the number of false negatives over the set of this category. In contrast, FB-IoU [25] ignores the difference among the object categories and regards all the categories as foreground class, and then averages the IoU of foreground and background over all test images.

### 4.1 PASCAL-$5^i$

The PASCAL-$5^i$ dataset combines images from the PASCAL VOC 2012 [5] and extra annotations from SDS [7]. We follow the dataset division in [28], in which the original 20 object classes in the official resealed order of PASCAL VOC are

**Table 1.** Performance comparison on PASCAL-$5^i$.

| Methods | Mean-IoU(1-shot) | | | | | FB-IoU | Mean-IoU(5-shot) | | | | | FB-IoU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | s-0 | s-1 | s-2 | s-3 | mean | (1-shot) | s-0 | s-1 | s-2 | s-3 | mean | (5-shot) |
| OSLSM [28] | 33.6 | 55.3 | 40.9 | 33.5 | 40.8 | 61.3 | 35.9 | 58.1 | 42.7 | 39.1 | 43.9 | 61.5 |
| co-FCN [25] | 36.7 | 50.6 | 44.9 | 32.4 | 41.1 | 60.1 | 37.5 | 50.0 | 44.1 | 33.9 | 41.4 | 60.2 |
| AMP-2 [29] | 41.9 | 50.2 | 46.7 | 34.7 | 43.4 | 61.9 | 40.3 | 55.3 | 49.9 | 40.1 | 46.4 | 62.1 |
| SG-One [41] | 40.2 | 58.4 | 48.4 | 38.4 | 46.3 | 63.1 | 41.9 | 58.6 | 48.6 | 39.4 | 47.1 | 65.9 |
| PANet [36] | 42.3 | 58.0 | 51.1 | 41.2 | 48.1 | 66.5 | 51.8 | 64.6 | 59.8 | 46.5 | 55.7 | 70.7 |
| CANet [40] | 52.5 | 65.9 | 51.3 | 51.9 | 55.4 | 66.2 | 55.5 | 67.8 | 51.9 | 53.2 | 57.1 | 69.6 |
| PGNet [39] | **56.0** | 66.9 | 50.6 | 50.4 | 56.0 | 69.9 | 57.7 | 68.7 | 52.9 | 54.6 | 58.5 | 70.5 |
| FWB [22] | 51.3 | 64.5 | 56.7 | **52.2** | 56.2 | - | 54.8 | 67.4 | **62.2** | **55.3** | 59.9 | - |
| **DAN** (Ours) | 54.7 | **68.6** | **57.8** | 51.6 | **58.2** | **71.9** | **57.9** | **69.0** | 60.1 | 54.9 | **60.5** | **72.3** |

evenly divided into four folds and conduct cross-validation over all the folds. Specifically, 15 object categories are used during training while the remaining 5 are used for testing for each fold. Following the settings in [28], we use 1000 support-query pairs of test support-query images for each test class.

We compare our proposed DAN with the state-of-the-art methods. As shown in Table 1, our DAN significantly outperforms all previous models under both 1-shot and 5-shot settings by large margins. Specifically, in the 1-shot setting, DAN achieves 58.2% and 71.9% in terms of Mean-IoU and FB-IoU, respectively. DAN outperforms the state-of-the-art graph-based method [39] by 2.2% in the Mean-IoU metric and 2.0% in the FB-IoU metric, which demonstrates the great benefit of our democratized graph attention mechanism. A comparison of 5-shot results is shown in Table 1, where we can see that our DAN achieves the highest performance under both evaluation metrics. Fig. 4 shows some qualitative segmentation maps produced by the DAN. As can be seen, our DAN can produce accurate segmentation maps in challenging cases where the objects in the query images show great variations in both size and appearance from the annotated objects in the support images.

### 4.2 COCO-$20^i$

The COCO-$20^i$ dataset is created for evaluation from a more challenging dataset MSCOCO [18]. The scenes in MSCOCO are more complex and the number of the categories is much higher than PASCAL-$5^i$. Similarly to Pascal-$5^i$, the 80 object categories in MSCOCO are evenly divided into four splits for cross-validation. For each split, 20 classes are sampled for testing and the remaining 60 classes are utilized for training. 1000 support-query pairs of support-query images are sampled from the 20 test classes for testing on each split.

The comparison results with previous methods are reported in Table 2. In the 1-shot setting, our DAN produces a new state-of-the-art performance, which is significantly better than the previous methods by 3.5% and 3.1% in terms of Mean-IoU and FB-IoU, respectively. The performance improvement obtained by our DAN demonstrate its great capability of handling complex scenes. In the 5-shot setting, our DAN also achieves comparable performance with the state-of-the-art method [36]. We show some qualitative visualizations in Fig. S1. Our DAN can perform very well in challenging cases. For instance, in the first and seventh cases, the objects in the query images are much smaller than those in their support images, while our DAN is still able to predict accurate segmentation maps; in the second case, the appearance of the object in query image is significantly different from that in the support image, but our DAN can still produce a segmentation result close to the ground truth.

### 4.3 FSS-1000

The FSS-1000 dataset contains 1000 object classes, a significant number of which have never been seen in previous datasets. Following [38], the 1000 categories

are divided into three splits for training, validation and testing. The training/validation/test splits consist of 520/240/240 classes, respectively. There are only 10 support-query pairs in each category. Note that the metric on FSS-1000 is the Intersection-over-Union of positive labels (P-IoU) in binary segmentation maps as in [38]. The comparison results are shown in Table 3. As can be seen, our DAN achieves the best performance in both 1-shot and 5-shot settings, with P-IoUs of 85.2% and 88.1%, respectively, largely surpassing the previous best performance. The improvement is over 10% in the 1-shot setting, showing the great performance advantage.

**Table 2.** Performance comparison on COCO-$20^i$.

| Methods | 1-shot | | 5-shot | |
|---|---|---|---|---|
| | Mean | FB | Mean | FB |
| A-MCG [11] | - | 52.0 | - | 54.7 |
| FWB [22] | 21.2 | - | 23.7 | - |
| PANet [36] | 20.9 | 59.2 | **29.7** | 63.5 |
| **DAN** (Ours) | **24.4** | **62.3** | 29.6 | **63.9** |

**Table 3.** Performance comparison on FSS-1000.

| Methods | P-IoU | |
|---|---|---|
| | 1-shot | 5-shot |
| OSLSM [28] | 70.3 | 73.0 |
| GNet [24] | 71.9 | 74.3 |
| FSS [38] | 73.5 | 80.1 |
| **DAN** (Ours) | **85.2** | **88.1** |

### 4.4 Ablation Study

We conduct extensive ablation experiments on PASCAL-$5^i$, results of which are reported as the average of the four folds under the Mean-IoU metric.

**Backbone network.** To evaluate the influence of different backbones, we experiment with two backbone models used in previous works: ResNet-50 [39, 40] and ResNet-101 [22]. The Table 4 shows the results of DAN with ResNet-50 and ResNet-101 on PASCAL-$5^i$.

**Table 4.** Results of our DAN with different backbones on PASCAL-$5^i$.

| Backbone | Mean-IoU | |
|---|---|---|
| | 1-shot | 5-shot |
| ResNet-50 | 57.1 | 59.5 |
| ResNet-101 | 58.2 | 60.5 |

**Table 5.** Influence of multi-scale evaluation on PASCAL-$5^i$.

| Model | Mean-IoU | |
|---|---|---|
| | 1-shot | 5-shot |
| DAN w/o MS | 57.5 | 60.1 |
| DAN w/ MS | 58.2 | 60.5 |

**Benefit of democratized graph attention.** To illustrate the improvement brought by the Democratized Graph Attention, we implement a baseline model with the regular graph attention (GA) mechanism used in PGNet [39]. We progressively mask the support images by different proportions, as shown in Fig 6, and evaluate the Mean-IoU performance under different ratios of invisible support images. Fig. 5 shows a comparison between the models with DGA and GA under different ratios of occlusion, from 0 to 0.8. Note that the result of ratio 0 is on original PASCAL-$5^i$. The Mean-IoU of the GA model drops rapidly from 56.6% to 41.7% in the 1-shot setting, while our proposed DGA shows relatively

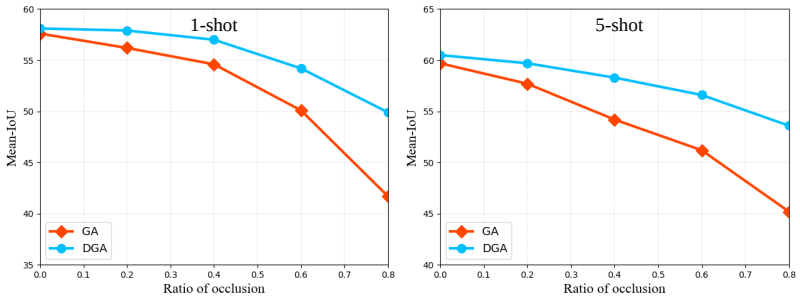



**Fig. 5.** Performance under different ratios of occlusion. The proposed DGA performs much better than the regular GA under occlusion.

stable performance, dropping from just 58.1% to 49.9%. Particularly, when the occlusion rate reaches 0.8, DGA outperforms the GA with a large margin of 8.2%. This indicates the vital role of our democratized graph attention mechanism in establishing robust connections between pixels of the support and query images, resulting in more robust segmentation maps.

We show the segmentation maps with different ratios of occlusion in Fig. 6. As can be seen, the activated regions in the attention map provided by the proposed DGA are much larger than those provided by the regular GA. This makes our DGA more robust to the partial occlusion of support objects. With the increase of occlusion ratios, the performance of GA declines rapidly, while the DGA is not affected much. In particular, the regular GA misses segmenting some of the foreground objects (left) or fails to segment the whole object (right), while our DGA is still able to successfully segment all the objects (left) and most parts of the object (right).

**Benefit of multi-scale guidance.** To demonstrate the benefit of multi-scale guidance, we conduct experiments on our DAN in which we gradually remove

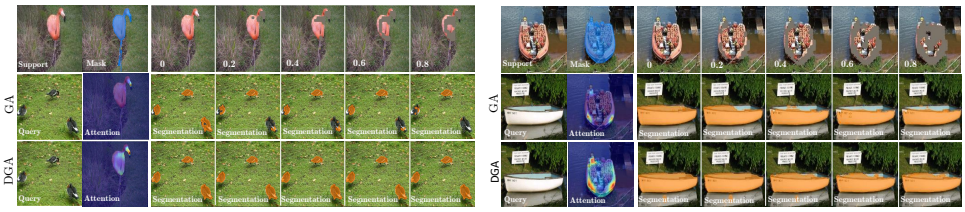



**Fig. 6.** One-shot segmentation maps under different ratios of occlusion. The DGA is more robust to occlusion than the regular GA.

**Table 6.** Benefit of multi-scale guidance. The performance is largely improved by the proposed multi-scale guidance, especially in the 1-shot setting.

| level 1 | level 2 | level 3 | Mean-IoU | |
|---|---|---|---|---|
| | | | 1-shot | 5-shot |
| | | ✓ | 56.4 | 59.2 |
| | ✓ | ✓ | 57.8 | 60.2 |
| ✓ | | ✓ | 56.9 | 59.7 |
| ✓ | ✓ | ✓ | **58.2** | **60.5** |

different levels of guidance. In Table 6, we compare the performance of the variants of the models with different levels of guidance. We can see that the more levels of information used, the better the performance. The performance reaches its highest when all levels are used. It is worth mentioning that multi-scale guidance offers more benefit in the 1-shot than 5-shot setting. This is because, in the 1-shot setting, we have only one support image, from which our multi-scale guidance can extract more guiding information for segmentation.

**Multi-scale inference.** We test the performance with multi-scale inference, which is adopted in [40] and [39]. Specifically, we rescale the query image by $\{0.5, 1, 1.5\}$ and average the predicted results. As shown in Table 5, multi-scale inference brings 0.7% and 0.4% improvements in 1-shot and 5-shot settings, respectively.

## 5 Conclusion

In this paper, we propose a new, democratic attention network (DAN) for few-shot semantic segmentation. We introduce a democratized graph attention mechanism to endow the network with the ability to establish robust connection between support and query images. Furthermore, we propose a multi-scale guidance structure by exploiting multiple levels of semantic information from the support images to guide the segmentation of query images. Extensive experiments on three benchmark datasets show that our DAN significantly outperforms previous works and achieves a new state-of-the-art performance. Thorough ablation studies demonstrate the benefits of the proposed democratized attention mechanism and multi-scale guidance for few-shot semantic segmentation.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under grant NO. 91738301, 61871016, and the National Key Scientific Instrument and Equipment Development Project under Grant NO. 61827901.

## References


1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on Pattern Analysis and Machine Intelligence **39**(12), 2481–2495 (2017)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on Pattern Analysis and Machine Intelligence **40**(4), 834–848 (2017)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009)
4. Dong, N., Xing, E.: Few-shot semantic segmentation with prototype learning. In: British Machine Vision Conference. vol. 1, p. 6 (2018)
5. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2), 303–338 (2010)
6. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1126–1135. JMLR. org (2017)
7. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: Proceedings of the International Conference on Computer Vision. pp. 991–998. IEEE (2011)
8. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969 (2017)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016)
11. Hu, T., Yang, P., Zhang, C., Yu, G., Mu, Y., Snoek, C.G.: Attention-based multi-context guiding for few-shot semantic segmentation (2019)
12. Hu, Y., Yang, Y., Zhang, J., Cao, X., Zhen, X.: Attentional kernel encoding networks for fine-grained visual categorization. IEEE Transactions on Circuits and Systems for Video Technology (2020)
13. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2014)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105 (2012)
15. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1925–1934 (2017)
16. Lin, G., Shen, C., Van Den Hengel, A., Reid, I.: Efficient piecewise training of deep structured models for semantic segmentation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 3194–3203 (2016)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755. Springer (2014)

18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755 (2014)
19. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015)
20. Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F.: See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In: CVPR. pp. 3623–3632 (2019)
21. Lu, X., Wang, W., Shen, J., Tai, Y.W., Crandall, D.J., Hoi, S.C.: Learning video object segmentation from unlabeled videos. In: CVPR. pp. 8960–8970 (2020)
22. Nguyen, K., Todorovic, S.: Feature weighting and boosting for few-shot segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 622–631 (2019)
23. Nichol, A., Schulman, J.: Reptile: a scalable metalearning algorithm. arXiv preprint arXiv:1803.02999 **2**, 2 (2018)
24. Rakelly, K., Shelhamer, E., Darrell, T., Efros, A.A., Levine, S.: Few-shot segmentation propagation with guided networks. arXiv preprint arXiv:1806.07373 (2018)
25. Rakelly, K., Shelhamer, E., Darrell, T., Efros, A., Levine, S.: Conditional networks for few-shot semantic segmentation (2018)
26. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: International Conference on Learning Representations (2017)
27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
28. Shaban, A., Bansal, S., Liu, Z., Essa, I., Boots, B.: One-shot learning for semantic segmentation. In: British Machine Vision Conference (2017)
29. Siam, M., Oreshkin, B.N., Jagersand, M.: Amp: Adaptive masked proxies for few-shot segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5249–5258 (2019)
30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
31. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. pp. 4077–4087 (2017)
32. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**, 1929–1958 (2014)
33. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1199–1208 (2018)
34. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: Proceedings of the International Conference on Learning Representations (2018)
35. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks one shot learning. In: Advances in Neural Information Processing Systems. pp. 3630–3638 (2016)
36. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9197–9206 (2019)

37. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 7794–7803 (2018)
38. Wei, T., Li, X., Chen, Y.P., Tai, Y.W., Tang, C.K.: Fss-1000: A 1000-class dataset for few-shot segmentation. arXiv preprint arXiv:1907.12347 (2019)
39. Zhang, C., Lin, G., Liu, F., Guo, J., Wu, Q., Yao, R.: Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9587–9595 (2019)
40. Zhang, C., Lin, G., Liu, F., Yao, R., Shen, C.: Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 5217–5226 (2019)
41. Zhang, X., Wei, Y., Yang, Y., Huang, T.: Sg-one: Similarity guidance network for one-shot semantic segmentation. arXiv preprint arXiv:1810.09091 (2018)
42. Zhen, X., Sun, H., Du, Y., Xu, J., Yin, Y., Shao, L., Snoek, C.: Learning to learn kernels with variational random features. International Conference on Machine Learning (2020)