International Journal of Computational Intelligence Systems
Vol. **14(1)**, 2021, *pp.* 886–895
DOI: https://doi.org/10.2991/ijcis.d.210212.003; ISSN: 1875-6891; eISSN: 1875-6883
https://www.atlantis-press.com/journals/ijcis/

Research Article

Few-Shot Image Segmentation Based on Dual Comparison Module and Sequential k-Shot Integration

Chencong Xing[1,*], Shujing Lyu[2], Yue Lu[2]

[1]*School of Computer Science and Technology, East China Normal University, Shanghai, 200241, China*
[2]*Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai, 200241, China*

ARTICLE INFO

Article History
Received 12 Jan 2021
Accepted 03 Feb 2021

Keywords
Few-shot learning
Image segmentation
Dual comparison module
Convolutional-gated recurrent unit

ABSTRACT

Few-shot image segmentation intends to segment query images (test images) given only a few support samples with annotations. However, previous works ignore the impact of the object scales, especially in the support images. Meanwhile, current models only work on images with the similar size of the object and rarely test on other domains. This paper proposes a new few-shot segmentation model named DCNet, which fully exploits the support set images and their annotations and is able to generalize to the test images with unseen objects of various scales. The idea is to gradually compare the features from the query and the support image, and refine the features for the query. Furthermore, a sequential k-shot comparison method is proposed based on the ConvGRU to integrate features from multiple annotated support images. Experiments on Pascal VOC dataset and X-ray Security Images demonstrate the excellent generalization performance of our model.

1. INTRODUCTION

Image segmentation is one of the basic tasks in the computer vision. Deep neural networks have significantly promoted its development in recent years. By fitting on large-scale datasets which have thousands of annotated images, such as PASCAL VOC [1] and MS COCO [2], deep learning models gain the ability to segment the objects. However, pixel-level labels are extremely costly in some situations, and current models can barely segment object from new categories after trained on pre-defined classes without the extra finetuning. In contrast, humans can segment new categories of objects easily without seeing too much samples, which implies the significant potential improvement for deep neural networks.

Few-shot image segmentation, which aims to predict unseen categories with a few annotated images after training, is critical for real applications. Compared with the common segmentation approaches [3–6], few-shot segmentation focuses on mining the correlation between the query images and the support images, which greatly improves the generalization ability given the limited annotations.

Few-shot segmentation is a new research area. Taking a pair of images as input, the approach of few-shot segmentation is able to segment the query image guided by the similarity between the annotated support images and the query image. Shaban *et al.* [7] first analyzed the few-shot semantic segmentation problem and proposed a two-branch model named OSLSM to process support images and query images separately. Weight hashing operation was designed to mix semantic context and generate the parameters for the query branch. Rakelly *et al.* [8] proposed the guided network, which extracted a latent representation from annotated pixels including positive and negative pixels, and further made the comparison by concatenating the query features and support features. Meanwhile, Dong and Xing [9] inherited the same framework of OSLSM and designed a prototype learner to extract features and generate the parameters for the query branch. Zhang *et al.* [10] designed a dense comparison module to compute the similarity by simply concatenating query and support features, and an iterative optimization module to gradually refine the results. They used global average pooling (GAP) over the support features to eliminate irrelevant information.

For few-shot task, most of the models in previous methods applied 1-shot method independently to each support example and use simple fusion methods to fuse individual results at the image level or feature level. Shaban *et al.* [7] proposed to use logic OR operation to fuse individual masks. Rakelly *et al.* [8] averaged the features in the support branch generated by different support samples. Zhang *et al.* [10] proposed an attention mechanism for k-shot task and used the softmax function to normalize the outputs of the attention module from different support samples.

There are some common limitations among these methods which remain to be solved: 1) The size of the objects in support images can seriously affect the segmentation results. Some operations, such as GAP or weight hashing, are detrimental for multi-scale targets, because these operations ignore the relevant spatial region of the

*Corresponding author. Email: 51184506047@stu.ecnu.edu.cn

object. 2) In k-shot settings, previous methods mainly use non-learnable methods or attention mechanism to directly fuse features from k-shot support images, which is inadequate according to the poor results. 3) Current few-shot segmentation models are only tested on natural images, which is not enough to prove the generalization ability of the model. In other image domains such as X-ray images which has the variety of object scales, viewpoints and heavy occlusions, few-shot segmentation may have a performance gap.

To overcome these problems, we propose the dual comparison network named DCNet for the task of few-shot segmentation. Figure 1 shows the overview of the proposed network. Our model measures the similarity between support images and query images from different perspectives on different scales. Since objects in the query images and support images may have different viewpoints and scales, we design the dual comparison module, which can be divided into the fine comparison submodule and the coarse comparison submodule, to deal with these problems. The fine comparison submodule utilizes the modified nonlocal operation [11] to capture the pixel-level similarity efficiently. In the coarse comparison submodule, we utilize the mask average pooling (MAP) [12] to get a representation of annotated areas while avoiding the impact of object scales. Furthermore, we present an operation based on ConvGRU to extend the 1-shot model to k-shot ways. Features from different support images are viewed as a sequence, the ConvGRU operation gradually fuses the feature sequence and produces the synthetic support features for the dual comparison module.

We conduct comprehensive experiments on the Pascal-5^i dataset to verify the performance, and also on our X-ray prohibited item segmentation dataset to prove the generalization. Experiments illustrate that our model achieves significant performance on few-shot segmentation and has good generalization ability to different data domains. Specifically, our model achieves 58.2% and 60.1% under the mean-IoU metric for the 1-shot task and 5-shot on Pascal-5^i, respectively.

Main contributions of this paper are summarized as follows:

- We propose a new few-shot segmentation model named DCNet. The dual comparison module is the core part of the model. The fine comparison submodule of dual comparison module is based on the nonlocal operations and is capable of measuring the dense spatial similarity between the pair of features for the query images and the support images. In the coarse comparison submodule, we utilize the MAP to focus on the relevant areas in the support image and get a more precise representation compared with other operations of pooling.

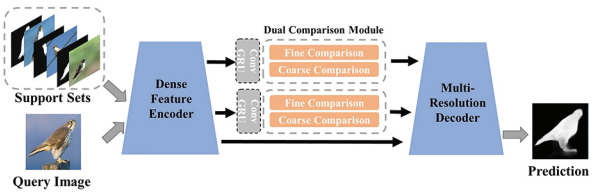


Figure 1 Overview of our proposed DCNet for the task of few-shot segmentation. Our DCNet can give results for new categories with at least one annotated samples.

- We propose a sequential k-shot method base on the ConvGRU, which can integrate the features from k-shot support images of a specific category by keeping in mind the global context in a sequential way.
- The proposed approach achieves significant performance gains from the state-of-the-art (58.2% vs. 56.0% and 60.1% vs. 58.5% on the Pascal-5^i dataset for 1-shot task and 5-shot task, respectively). Meanwhile, the experiment results on our X-ray image dataset also demonstrate the generalization on different image domain.

2. RELATED WORKS

Weakly Supervised Segmentation aims to find a feasible method for semantic segmentation to reduce the cost of pixel-level annotation. Hong *et al.* [13] proposed a weak supervised model which combined the classification and segmentation into one network. With only image-level labels, the proposed model can learn to segment according to the class activation map (CAM) [14]. Zhang *et al.* [15,16] took inspiration from the CAM and proposed to use classification network to discover the regions of objects. Lin *et al.* [17] applied spectral clustering method to classify the object pixels according to the similarity of adjacent pixels and ground-truth scribble lines. Wei *et al.* [18,19] proposed a two-stage architecture to generate segmentation masks. First, the coarse regions of objects were determined by the CAM. The segmentation part, such as DeepLab [20], then trained to produce the segmentation masks.

Few-Shot Learning aims to learn transferable knowledge that can be generalized to unseen categories with only a few annotated images. There exist some solutions on few-shot classification, including learning to finetune models [21], meta-learning [22,23] and metric learning [24]. Finetune-based methods [21] utilize the limited samples to refine the model which have trained on other large-scale dataset. Meta learning-based approaches address the few-shot learning problem by training networks to learn how to learn novel classes. Ravi and Larochelle [22] focused on the similarity between gradient descent methods and long–short-term memory (LSTM). They achieved a fast adaptation to unseen classes by using LSTM to update network weights. Metric learning-based methods achieve state-of-the-art performance in the few-shot classification tasks, and they have the trait of being fast and predicting in a feed-forward manner. Relation Network [24] learned an abstract distance metric to compare images and computed the similarity score for classification. The network consisted of an encoder module which generates the representations of the images and a relation module that compares the image features and outputs a similarity score. The dual comparison module in our model can be seen as an extension of Relation Network in a dense form to tackle the task of segmentation.

3. METHODS

In this section, we first give a problem formulation of few-shot segmentation task, and then introduce the proposed model for 1-shot task in detail. Finally, we show the method to extend our 1-shot segmentation model to few-shot in the sequential way.

3.1. Problem Definition

Suppose we have a label set $C = \{c_i\}_{i=1}^{N_c}$, where c_i represents the ith specific category, such as *car, bus*.

N_c is the total number of categories. Based on C, there is an annotated support set $S = \{(I_{si}^{c_x}, L_{si}^{c_x})\}_{si=1}^{N_s}$, where $I_{si}^{c_x}$ is the ith image in the support set and it belongs to category c_x, $L_{si}^{c_x} i$ is the binary mask for $I_{si}^{c_x}$; a query set $Q = \left\{ I_{qi}^{c_y} \right\}_{qi=1}^{N_q}$, where $I_{qi}^{c_y}$ is the ith image in the query set, the category is c_y.

The task of few-shot segmentation is that given an unseen query image $I_q^{c_m}$ and k support image-label pairs $\{(I_{si}^{c_m}, L_s i^{c_m})\}_{si=1}^{k}$, the model should have the ability to find out relevant areas in the query images and output the predicted mask $L_q^{c_m}$. It is generally accepted that k should be equal to or larger than 5 in few-shot segmentation task. When k=1, few-shot segmentation is equivalent to 1-shot segmentation. Moreover, to train or measure the segmentation model, the label set is split into C_{train}, C_{test}. We ensure that $C_{train} \cap C_{test} = \varnothing$. The support set can also be divided into S_{train} and S_{test} according to the categories. There is no duplicated image-label pairs appearing in both sets. Few-shot segmentation model trained on S_{train} and measure on the S_{test}.

3.2. Proposed Model

In this paper, we design a dual comparison network for the task of few-shot segmentation. The proposed model first adopts a dense feature encoder to extract the query and support features simultaneously. Dual comparison modules are implemented in the middle two resolutions (block 3 and block 5). Inspired by the non-local operation [11,25], the fine comparison submodule pays attention to the spatial element-wise similarity between the query images and support images. We design the architecture to compute the dense similarity by a series of matrix operations. In the part of the coarse comparison submodule, we utilize the MAP to eliminate the irrelevant information and get the precise representations of support features. We compare the query features and support features by concatenation to measure the regional similarity. Finally, the multi-resolution decoder integrates query features with different resolutions gradually and further generates the segmentation mask. The following part gives a detailed introduction of our model. Figure 2 shows the architecture of our proposed model.

Dense Feature Encoder indicates to extract feature representations at different levels from the query and support images simultaneously. Convolutional neural network is the most widely used feature extractor. As is observed in the visualization of the convolutional neural networks and previous few-shot segmentation research [10], features in the first several layers contain more information about the edges and colors than the last few layers which have stronger class information, while middle layers have more attribute information which is shared by different categories. Since the training set and test set have no duplicate categories, we cannot assume that the feature encoder can extract the features of unseen categories during training. Middle-level attribute features are more important to the few-shot task. Therefore, we choose the middle-level features from block 3 and 5 of the dense feature encoder for the dual comparison modules. Both the support and query branch use the same feature encoder.

In this paper, we use a feature extractor backbone on the basis of dilated residual network (DRN) [26]. Considering the impact of the object scales, we modified the DRN-C-42 model. Specifically, as is shown in Figure 3, we add deformable convolution [27] to the end of block 3, block 4, and block 5 to make the encoder module gain the ability to extract the feature of objects with different sizes.

Dual Comparison Module is the core part of our proposed model. It can be divided into two independent submodules: the fine comparison submodule and the coarse comparison submodule. The similarity calculated in both comparison submodules can be viewed as the supervision information, which guide the decoder to generate the segmentation mask.

Fine Comparison Submodule intends to compare the query and support features to capture the spatial element-wise similarity between all pairs of pixels. The objects in the query images and the support images may have various viewpoints, scales, and even overlapping conditions. Our proposed fine comparison submodule has the ability to compare the features densely by extending the nonlocal operations. The fine comparison operation can be defined as Eq. (1):

$$\begin{cases} M_{sm} = Softmax\left[\theta\left(x_{qry}\right)^{\top} \otimes \left(m \cdot \varphi\left(x_{sup}\right)\right)\right] \\ Fine_{\ out} = x_{qry} \otimes M_{sm} \end{cases} \tag{1}$$

where x_{qry} and x_{sup} are middle-level features from dense feature encoder. They contain middle-level semantic context which shared by all kinds of categories including unseen classes. $\theta(\cdot)$ and $\phi(\cdot)$ represent linear transformations which are implemented as 1×1 convolutions to half the number of channels on both support and query features for reducing redundant information. To measure the dense similarity between x_{qry} and x_{sup}, we first filter out the irrelevant areas on support features by multiply the support mask m. Then both query features and support features are reshaped from $(N,C/2,H,W)$ to $(N,C/2,HW)$. We compute the matrix production of them to get the similarity matrix (M_{sm}). Each row of the similarity matrix represents the similarity of one pixel in the query feature map to all pixels in the support feature map. We use softmax function to normalize this similarity matrix line by line. The output of the fine comparison submodule is the matrix production of x_{qry} and M_{sm}.

From the spatial perspective, our fine comparison submodule has the ability to capture the pixel-level similarity despite the relative distance of two pixels. This methods is more precise and interpretable than previous works in few-shot segmentation task. Similarity matrix can give not only dense feature similarity but also the evidence of how the model transferring the annotated information from support images to query images.

Coarse Comparison Submodule focuses on the regional similarity between query images and support images. Since they are encoded by the same backbone, similar objects must share some common patterns in the feature maps. Compared with element-wise similarity, regional similarity can filter out noise interference and produce region-consistent feature correspondence.

A simple method of the coarse comparison is to adopt GAP [28] to squeeze support features and get global representation of the object. The squeezed support features are further replicated to the same shape as the query features and connected with the query feature along the axis of channels. We adopt the similar method. However, GAP is detrimental for dealing with multi-scale targets because

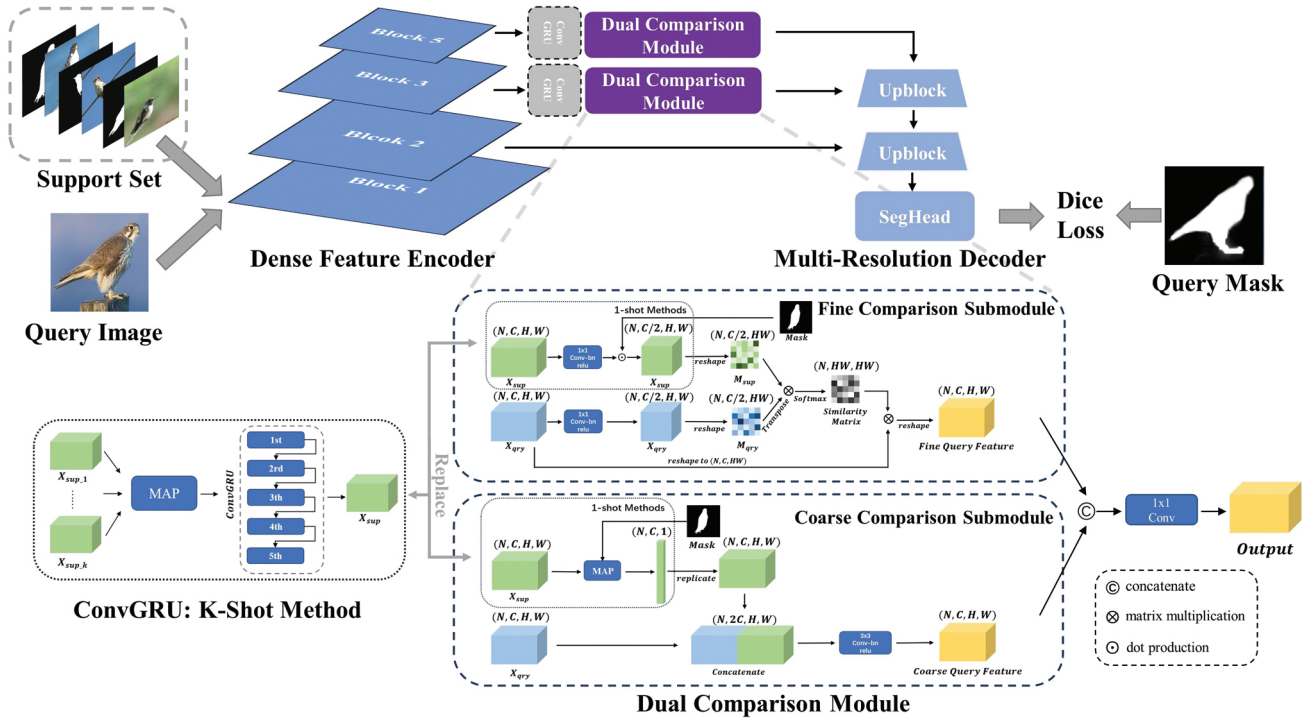


Figure 2 The architecture of the proposed DCNet in detail. After training, the proposed model can recognize novel categories with only a few annotated samples.

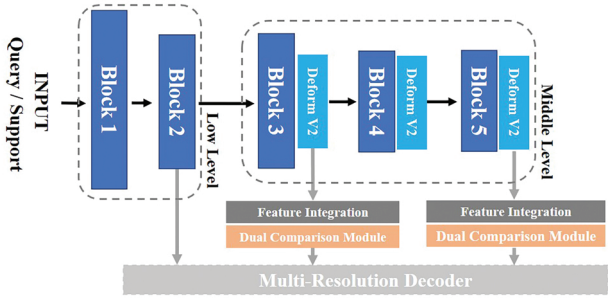


Figure 3 The modified dense feature encoder in detail. The deformable convolution has been added to the end of block 3, block 4, and block 5.

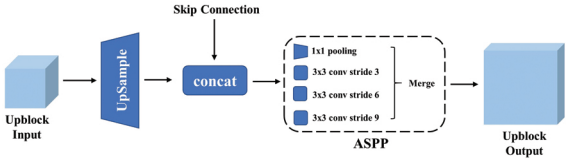


Figure 4 Upblock in the multi-resolution decoder. Upblock makes fusion of features with different resolution. We also introduce the ASPP module to adapt to different scales of the targets.

objects will take a different percentage of feature vectors after GAP, which causes the model to have apparent performance bias.

On the basis of GAP, we utilize the MAP [12], which can take the annotation information to force the pooling operation on the target area to retain more useful information. The calculation of MAP can be defined as:

$$MAP = \frac{\sum_{\forall i} x_i \cdot 1\left\{ Mask_{\,i} = 1 \right\}}{\sum_{\forall i} 1\left\{ Mask_{\,i} = 1 \right\}} \tag{2}$$

where i indexes the spatial locations and $1\{\cdot\}$ is an indicator function that outputs value 1 if the argument is true, 0 for otherwise. Compared with the GAP, MAP can achieve a more sustainable response with different scales targets. In the coarse comparison submodule, we concatenate the query feature and the replicated support representation after MAP along the axis of channels to make regional comparison. We also add a 3x3 conv-bn-relu block to further integrate the features and squeeze the channels.

Dual comparison module intends to fuse query and support information from the pixel-level and regional-level perspectives, respectively. In order to compute the result of the dual comparison module, we simply concatenate the fine comparison features and coarse comparison features, and use 1x1 convolution to adjust the fusion feature and squeeze the number of channels.

Multi-Resolution Decoder gradually recovers semantic information which has lost during down-sampling in the dense feature encoder. This module mainly consists of two upblocks and a segmentation head. As illustrated in Figure 4, the two up-blocks aims at upsampling the input feature and making the fusion of features with different resolutions. ASPP layer [20] is also used to adapt to target with different scales. The segmentation head is the final submodule of the multi-resolution decoder, which squeezes features and generates the binary mask. In this work, we use two ResBlocks to adjust the channels of the feature, and take the sigmoid function to acquire the final segmentation mask.

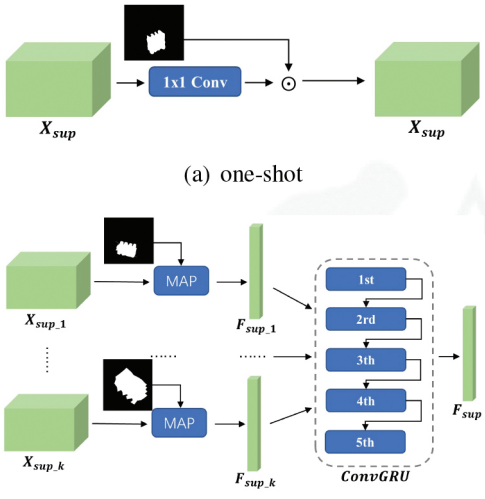
3.3. Sequential k-Shot Integration

In the case of k-shot segmentation, the support set contains k-annotated images. To extend the dual comparison module to cover the k-shot situation, we propose the methods of sequential k-shot integration based on ConvGRU separately for the fine comparison submodule and the coarse comparison submodule. Figure 5(b) shows the specific operations to extend the 1-shot method.

ConvGRU represents the convolutional-gated recurrent units [29], which are a gating mechanism in recurrent neural networks. Compared with the architecture of traditional LSTM [30], GRU integrates the forget gate with the input gate and has fewer parameters. However, common GRU network is designed for the sequence data and uses fully connection layers to process the temporal relation, which is not suitable for high-dimensional image data. Shi *et al.* [31] inspire by the ConvLSTM [32] and propose the ConvGRU by modifying the fully connection operation to convolutional operation, which can effectively extract the spatial features of images. In this paper, we select ConvGRU instead of ConvLSTM for the sequential k-shot integration since ConvGRU has a simple structure and only one hidden state which is more easier to extend to process the images.

In the method of sequential k-shot integration, features from different support images are viewed as the sequence data and are input to the ConvGRU module in turns to generate the integrated support features. We randomly select one image and take the feature as the initial hidden state of the ConvGRU. In contrast to the approach of concatenation or summation, the memory mechanism implemented by Con-vGRU allows global context of support features to be kept in memory while details are gradually being added. This architecture aims at imitating the mechanism allowing humans to focus on details of a specific category while keeping in mind its global appearance.

Fine comparison submodule focuses on the pixel-level similarity between the query feature and support feature. When given k-annotated images, we first extract support features separately and use 1×1 conv-bn-relu block to half the number of support features channels to reduce the computational complexity and multiply their own masks to remove the irrelevant areas. The support features are then input to the ConvGRU block to generate the integrated features. We can further process the integrated features following the methods mentioned in Section 3.2.



(a) one-shot

(b) few-shot

Figure 5 (a) The proposed one-shot method. (b) The proposed ConvGRU methods to extend 1-shot model to few-shot ways.

Coarse comparison submodule aims at the comparison of global representation of annotated support objects and the query features. To extend to the k-shot situation, we first conduct MAP on each support images and then deliver the features to ConvGRU block. The remaining steps can be processed according to the coarse comparison submodule described above.

4. EXPERIMENTS

We conduct comprehensive experiments on Pascal-5^i. To evaluate the performance and our X-ray prohibited item segmentation dataset to prove the generalization. Our network is a fully convolutional model, which can take images with different scales as inputs. To deal with the problem of multi-scale objects, dice loss [36] is adopted to guide and constrain the model during training. Our model is implemented with the PyTorch framework. We use SGD optimizer and train the model for 200 epochs on Nvidia 2080Ti GPUs. The learning rate is set to $2.5e^{-10}$ and weight decay is $1e^{-4}$. We choose a mini-batch of 5 on Pascal-5^i dataset and 1 on the X-ray image dataset. Moreover, we recommend to use GroupNorm [37], instead of BatchNorm to get better results when the batch size is smaller than 15.

We choose two evaluation metric methods: meanIoU and FB-IoU [8,10] for our experiments. The Intersection over Union (IoU) metric, also referred to as the Jaccard index, is essentially a method to quantify the percent overlap between the target mask and our prediction mask. The IoU metric measures the number of common pixels between the target mask and prediction mask divided by the total number of pixels present across both masks. Mean-IoU is the average foreground-IoU of test categories, which is widely used to evaluate the performance of segmentation models. We also use FB-IoU metric, which takes the background-IoU into account, to get a more comprehensive metric. Eq. (3) gives the specific calculation formula of foreground-IoU and background-IoU.

$$\begin{cases} foreground - IoU = \dfrac{mask \cap pred}{mask \cup pred}, \\ background - IoU = \dfrac{\neg\, mask \cap \neg\, pred}{\neg\, mask \cup \neg\, pred} \end{cases} \tag{3}$$

4.1. Pascal VOC Dataset

Pascal-5^i dataset is the most widely used dataset in the field of few-shot segmentation. It is first mentioned in OSLSM [7]. The Pascal-5^i dataset is built based on the entire PASCAL VOC 2012 dataset and extra annotation files from the SBD dataset [38]. The PASCAL VOC dataset contains 20 categories. Pascal-5^i has split them into 4 groups, and each group has 5 categories. To train on a specific split, we use all other splits as the training set, and the left split as the test set. In the training stage, two images are randomly selected from each class in the training set, one as the query image and the other as the support image (1-shot situation). In the test stage, we use the

Table 1 The result of 1-shot task and 5-shot task on Pascal-5^i dataset. Our model outperform all the previous methods under the metrics of meanIoU and FB-IoU. The best results are in bold.

Methods	1-shot						5-shot					
	split-0	split-1	split-2	split-3	meanIoU	FB-IoU	split-0	split-1	split-2	split-3	meanIoU	FB-IoU
Siamese [33]	28.1	39.9	31.8	25.8	31.4	57.6	/	/	/	/	/	/
OSVOS [34]	24.9	38.8	36.5	30.1	32.6	57.4	/	/	/	/	/	/
OSLSM [7]	33.6	55.3	40.9	33.5	40.8	61.3	35.9	58.1	42.7	33.9	43.9	61.5
GN [8]	36.7	50.6	40.9	32.4	41.1	60.1	37.5	50.0	44.1	39.4	41.4	60.2
CANet [10]	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 [35]	56.0	66.9	50.6	50.4	56.0	67.4	57.9	68.7	52.9	54.6	58.5	71.3
Ours	**53.9**	**69.7**	**55.4**	**53.2**	**58.2**	**72.1**	**54.9**	**69.9**	**57.3**	**58.3**	**60.1**	**74.3**

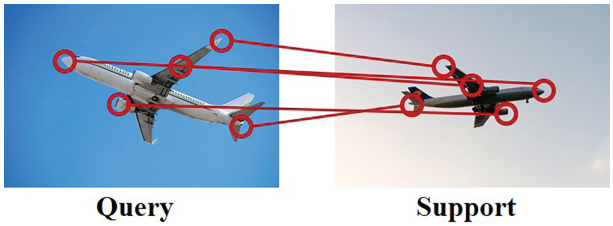


Figure 6 The visualization of similarity matrix in the fine comparison submodule. The red lines represent the pairs of feature points are relevant while the blue lines represent the wrong match for the pairs of feature points.

Table 2 The performance of CANet and the proposed DCNet with different backbones under 1-shot settings.

BackBone	Method	Pascal-5^i
VGG-16	CANet [10] DCNet	36.9 39.1
Resnet50	CANet [10] DCNet	55.4 56.8
DRN-C-42	CANet [10] DCNet	55.9 57.2
Our modified DRN	CANet [10] DCNet	56.1 58.2

same random seed to sample the same 1000 pairs of images as the test data to make a fair comparison with all othermethods.

4.1.1. Result and Comparison

We compared our method with other few-shot segmentation models on Pascal-5^i. Table 1 shows the results under the meanIoU and FB-IoU evaluation metrics. For the performance of 7, 8, 10, 33–35 under the FB-IoU metric, we quote the result reproduced in 10. Particularly, our model outperforms the state-of-the-art model by 2.2% and 1.6% for 1-shot and 5-shot, respectively. The FB-IoU score is 72.1% for 1-shot task and 74.3% for 5-shot task, which is slightly higher than the state-of-the-art model. In conclusion, our model performs to recognize the targets with high precision compared with all other approaches.

In Figure 6, we pick 5 highest responses from the similarity matrix of the fine comparison submodule and visualize the pixel-level relation according to the coordinate information. The red lines represent the pairs of feature points irrelevant while the blue lines represent the wrong match for the pairs of feature points. Figure 6 illustrates that fine comparison submodule is capable of capturing the fine-grained similarity.

Figure 7 shows the predicted results on Pascal-5^i. In Figure 7(a) and 7(b), we present the few-shot segmentation results, which prove that the proposed dual comparison model can segment objects with only a few annotated samples. We also show some failure examples on Pascal-5^i, see Figure 7(c). We think the differences in viewpoints and inter-class between the support images and the query images are part of the reasons for the failure.

4.1.2. Ablation study

We conduct ablation experiments on the Pascal-5^i dataset to illustrate the effectiveness of the proposed module in our DCNet. There are three modules for ablation studies: feature encoder backbone, dual comparison module and the sequential k-shot integration. The ablation experiments help us to confirm which module significantly contributes and determine whether these modules are necessary. We use meanIoU metric to evaluate the performance.

We compare different feature encoders to prove that our modified DRN backbone is better to our task. We choose VGG16 [39], Resnet50 [40], DRN42 [26] backbones to conduct the 1-shot segmentation and compare with our modified DRN network. Features with /4, /8 scales are extracted and transport to the dual comparison module for further processing. In Table 2, the results indicate that our modified DRN model is the best backbone compared with all others in this task. We believe that the dilated convolution with special dilation rates plays the critical role to make the trade-off between object features and location information.

To illustrate the ability of dual comparison module, we conduct experiments with 4 setups: without dual comparison module, only with fine comparison submodule (FCS), only with the coarse comparison submodule (CCS) and with the entire dual comparison module. As shown in Table 3, fine comparison submodule contributes 10.3% improvement of meanIoU compared without this submodule, which is less than 14.5% of the improvements achieved by coarse comparison submodule. The experiments indicate that dual comparison module play a critical role in the few-shot segmentation task.

Table 4 demonstrates the excellent performance of the sequential k-shot integration on Pascal-5^i dataset. We choose the attention mechanism proposed in the CANet as the baseline. For the ConvGRU method, we take two different order of input support images

(a) 1-shot result (b) 5-shot result (c) some failure examples

Figure 7 The result of our DCNet on Pascal-5^i dataset. (a) The 1-shot result. (b) The 5-shot result. (c) Some failure examples of 1-shot task. In each group of the results, the small image patches are the support samples, and images with red edge are the segmentation result.

Table 3 Ablation experiments of different settings for the dual comparison module on the Pascal-5^i dataset.

PCS	RCS	Mean IoU (%)
✓		43.7
	✓	47.9
✓	✓	58.2
		33.4

Table 4 Ablation experiments on the choice of k-shot methods. We compare the attention mechanism in CANet with our proposed ConvGRU methods (ConvGRU-random for random permutation the order of input support images).

Extensible Methods	1-shot	5-shot	10-shot	15-shot
Attention [10]	58.2	59.5	64.7	66.9
ConvGRU	58.2	59.3	65.6	67.6
ConvGRU-random	58.2	60.4	65.3	67.1

to prove that the permutation has no appearent effect on the performance. According to the results, we find that the proposed sequential k-shot integration based on ConvGRU has better performance to fuses support features and can achieve more significant improvement when given more support features.

Table 5 The average scale of objects in different categories.

Category	Scales		
	Smll $\leqslant 32^2$	Middle $> 32^2$ & $\leqslant 128^2$	Large $> 128^2$
Battery	✓		
Bottle		✓	
Explosive			✓
Firearm		✓	
Knife		✓	
Scissors	✓		
Lighter	✓		

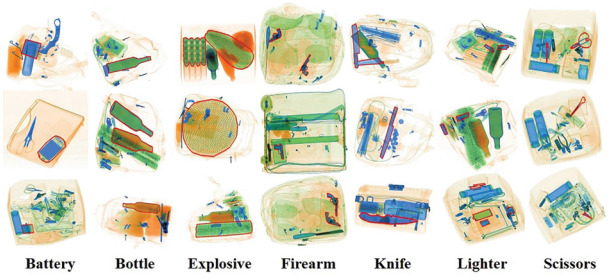


Figure 8 Some examples in the X-ray image dataset. The object sizes of different categories are quite different.

4.2. X-Ray Image Dataset

For a more comprehensive evaluation of model performance, we also conduct experiments on our X-ray security image segmentation dataset named X-ray-PI. X-ray security images have the significant variety in scales, viewpoints. We analyze the average scales of objects in X-ray images according to the standard of MS COCO dataset. As shown in Table 5, the object scales of different categories in X-ray-PI dataset are quite different. Moreover, compared with natural images, different colors in X-ray images represent different substances, and the problem of object occlusion also appears more frequently.

The X-ray-PI dataset contains 7 categories of prohibited items (Battery, Bottle, Explosive, Firearm, Knife, Lighter, Scissors) within total 2407 annotated images. Figure 8 shows some examples in the X-ray image dataset.

To experiment on our dataset, we select one category as the test class in turn and take all other categories as training classes. During the test, we randomly choose 1000 groups of query images and support images in each categories as inputs to quantify model performance.

Table 6 shows the results on the X-ray-PI dataset. All the models are fine tuned on the X-ray-PI dataset with pretrained weight from Pascal-5^i. Our model maintain the excellent performance and achieve the meanIoU of 51.5% and 55.8% for 1-shot and 5-shot, respectively. The results indicate that our model has better generalization performance and can adapt to different segmentation tasks.

Table 7 shows the results of objects with different scales on the X-ray-PI dataset. Our model outperforms the current few-shot segmentation models on all scales. However, the result of small objects still has a performance gap compared with the results of middle and large objects. We believe the downsam-pling operation such as maxpooling causes the irreversible information loss which has an obvious impact on small objects.

Figure 9 shows the predicted results on our X-ray image dataset. In Figure 9(a) and 9(b), we present the 1-shot and 5-shot segmentation

Table 6 the performance of different models on the X-ray-PI dataset.

Methods		Average Precision								FB-IoU
		Battery	Bottle	Explosive	Firearm	Knife	Scissors	Lighter	Avg.	
1-shot	GN [8]	45.0	45.3	40.6	40.6	24.2	31.6	32.8	37.2	52.9
	CANet [10]	47.7	62.1	61.4	67.0	20.6	21.4	40.8	45.8	62.1
	PGNet [35]	47.5	63.2	61.0	67.2	21.2	21.5	41.4	46.1	63.2
	Ours	**52.6**	**68.5**	**63.1**	**70.4**	**39.4**	**23.7**	**42.5**	**51.5**	**63.5**
5-shot	GN [8]	46.9	45.2	45.8	55.4	21.3	28.5	38.9	40.3	53.6
	CANet [10]	50.3	66.0	62.8	67.4	28.1	34.9	43.2	50.4	72.3
	PGNet [35]	50.7	67.2	63.5	68.3	29.2	34.6	43.5	51.0	73.9
	Ours	**55.4**	**73.7**	**67.3**	**73.6**	**50.2**	**25.5**	**44.9**	**55.8**	**75.2**

(a) 1-shot result (b) 5-shot result (c) some failure examples

Figure 9 The result of our DCNet on X-ray-PI dataset. (a) The 1-shot result. (b) The 5-shot result. (c) Some failure examples of 1-shot task. In each group of the results, the small image patches are the support samples, and images with red edge are the segmentation result.

Table 7 The performance of different object scales on the X-ray-PI dataset.

Methods		Average Precision				FB-IoU
		Small	Middle	Large	Avg.	
1-shot	GN [8]	29.5	40.9	41.2	37.2	52.9
	CANet [10]	40.7	47.1	49.6	45.8	62.1
	PGNet [35]	39.8	47.5	51.0	46.1	63.2
	Ours	**40.9**	**56.4**	**57.2**	**51.5**	**63.5**
5-shot	GN [8]	30.4	44.3	46.2	40.3	53.6
	CANet [10]	46.3	52.7	52.2	50.4	72.3
	PGNet [35]	41.6	57.9	53.5	51.0	73.9
	Ours	**50.0**	**57.6**	**59.8**	**55.8**	**75.2**

results, which prove that the proposed DCNet can segment images from differnet domain with only a few annotated samples. We also show some failure examples on X-ray-PI dataset, see Figure 9(c). We think the huge differences between natural images and X-ray images are part of the reasons for the failure.

5. CONCLUSION

In this paper, we propose the dual comparison network for the task of few-shot segmentation. The fine comparison submodule focuses on the spatial element-wise similarity while the coarse comparison submodule undertakes the regional similarity. By fusing the similarity of different scales, DCNet is capable of generating the precise results. Furthermore, a mechanism based on ConvGRU operation is designed to extend the 1-shot model to few-shot ways. Compared with the attentive methods, the ConvGRU operation turns out to be more effective. We conduct comprehensive experiments to illustrate the effectiveness of our approach. Our network sets a new benchmark for 58.2% and 60.1% meanIoU in 1-shot and 5-shot settings on Pascal-5^i dataset, respectively.

The future research mainly lies in two direction. First, we ignore the problem of different viewpoints. Different viewpoints of the same objects can produce significant divergence in the images. We are trying to add the information of viewpoints to networks and design a more efficient network. Second, the feature encoder backbone pretrained on natural images dataset may not suit other application scenarios such as X-ray security images. A new backbone especially designed to extract middle-level attribute features may further improve the performance.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

AUTHORS' CONTRIBUTIONS

All authors contributed to the work, and all authors read and approved the final manuscript.

ACKNOWLEDGMENTS

Thank reviewers and editors for their helpful comments on this paper.

REFERENCES

[1] M. Everingham, L.V. Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal Visual Object Classes (VOC) challenge, Int. J. Comput. Vison. 88 (2010), 303–338.

[2] L. Tsung-Yi, M. Michael, B. Serge, H. James, P. Pietro, R. Deva, D. Piotr, Z. Larry, Microsoft coco: common objects in context, in The European Conference on Computer Vision (ECCV), Zurich, Switzerland, 2014.

[3] J. Long, E. Shelhamer, T. Darrel, Fully convolutional networks for semantic segmentation, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015.

[4] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in The International Conference on Medical image computing and computer-assisted intervention (MICCAI), Munich, Germany, 2015.

[5] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 40 (2018), 834–848.

[6] M. Simfukwe, B. Peng, T. Li, Fusion of measures for image segmentation evaluation, Int. J. Comput. Intell. Syst. 12 (2019), 379–386.

[7] A. Shaban, S. Bansal, Z. Liu, I. Essa, B. Boots, One-shot learning for semantic segmentation, in The British Machine Vision Conference (BMVC), London, UK, 2017.

[8] K. Rakelly, E. Shelhamer, T. Darrell, A.A. Efros, S. Levine, Few-shot segmentation propagation with guided networks, in The IEEE International Conference on Learning Representations (ICLR) Workshop, Vancouver, Canada, 2018.

[9] N. Dong, E.P. Xing, Few-shot semantic segmentation with prototype learning, in The British Machine Vision Conference (BMVC), Newcastle, UK, 2018.

[10] C. Zhang, G. Lin, F. Liu, R. Yao, C. Shen, Canet: classagnostic segmentation networks with iterative refinement and attentive few-shot learning, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019.

[11] Z. Zhu, M. Xu, S. Bai, T. Huang, X. Bai, Asymmetric non-local neural networks for semantic segmentation, in The IEEE International Conference on Computer Vision (ICCV), Seoul South, Korea, 2019.

[12] X. Zhang, Y. Wei, Y. Yang, T. Huang, Sg-one: similarity guidance network for one-shot semantic segmentation, IEEE Trans. Cybern. 50 (2020), 3855–3865.

[13] S. Hong, H. Noh, B. Han, Decoupled deep neural network for semi supervised semantic segmentation, in Advances in Neural Information Processing (NIPS), Montréal CANADA, 2015.

[14] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016.

[15] X. Zhang, Y. Wei, J. Feng, Y. Yang, T. Huang, Adversarial complementary learning for weakly supervised object localization, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018.

[16] X. Zhang, Y. Wei, J. Feng, Y. Yang, T. Huang, Self-produced guidance for weakly-supervised object localization, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Munich, Germany, 2018.

[17] D. Lin, J. Dai, J. Jia, K. He, J. Sun, Scribblesup: scribble-supervised convolutional networks for semantic segmentation, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016.

[18] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, S. Yan, Object region mining with adversarial erasing: a simple classification to semantic segmentation approach, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017.

[19] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, S. Yan, Revisiting dilated convolution: a simple approach for weakly-and semi-supervised semantic segmentation, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018.

[20] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2016), 834–848.

[21] J. Li, K. Lu, Z. Huang, L. Zhu, H. Shen, Transfer independently together: a generalized framework for domain adaptation, IEEE Trans. Cybern. 49 (2019), 2144–2155.

[22] S. Ravi, H. Larochelle, Optimization as a model for few-shot learning, in International Conference on Learning Representations (ICLR), Toulon, France, 2017, pp. 11–18.

[23] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in Advances in Neural Information Processing (NIPS), Long Beach, USA, 2017, pp. 702–710.

[24] F.S.Y. Yang, T. Xiang, L. Zhang, P.H. Torr, T.M. Hospedales, Learning to compare: Relation network for few-shot learning, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 5119–5128.

[25] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019.

[26] F. Yu, V. Koltun, T. Funkhouser, Dilated residual networks, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017.

[27] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable convnets v2: more deformable, better results, in International Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 9300–9308.

[28] M. Lin, Q. Chen, S. Yan, Network in network, in The IEEE International Conference on Learning Representations (ICLR), Banff, Canada, 2014.

[29] C. Kyunghyun, V. M. Bart, G. Caglar, B. Dzmitry, B. Fethi, S. Holger, B. Yoshua, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in The IEEE Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014.

[30] H. Sepp, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997), 1735–1780.

[31] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, W. WOO, Deep learning for precipitation nowcasting: a benchmark and a new mode, in Advances in Neural Information Processing (NIPS), Long Beach, USA, 2017.

[32] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, W. WOO, Convolutional LSTM network: a machine learning approach for

precipitation nowcasting, in Advances in Neural Information Processing (NIPS), Montréal CANADA, 2015.

[33] C. Finn, P. Abbeel, E. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, International Conference on Machine Learning, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Sydney, Australia, (2017), 1126–1135.

[34] S. Caelles, K.-K. Maninis, L. LealTaixe, D. Cremers, L. Van Gool, One-shot video object segmentation, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017.

[35] C. Zhang, G. Lin, F. Liu, J. Guo, Q. Wu, R. Yao, Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seoul, South Korea, 2019.

[36] F. Milletari, N. Navab, S.A. Ahmadi, V-net: fully convolutional neural networks for volumetric medical image segmentation, in The IEEE Conference on 3D Vision (3DV), Stanford, CA, USA, 2016.

[37] Y. Wu, K. He, Group normalization[J], International Journal of Computer Vision, 128 (2020), pp. 742–755.

[38] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, J. Malik, Semantic contours from inverse detectors, in International Conference on Computer Vision (ICCV), Barcelona, Spain, 2011.

[39] S. Karen, Z. Andrew, Very deep convolutional networks for large-scale image recognition, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, USA, 2014.

[40] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2015.