Image Co-segmentation with Multi-Scale Dual-Cross Correlation Network

1st Yushuo Li
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China
liyushuo@bit.edu.cn

2nd Yuanpei Liu
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China
liuyuanpei@bit.edu.cn

3rd Xiaopeng Gong
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China
gongxp@bit.edu.cn

4th Xiabi Liu*
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China
liuxiabi@bit.edu.cn

***Abstract*—Considering that the global correlation between images is very important for image co-segmentation, we propose a multi-scale Dual-Cross Correlation Network (DCNet) that can efficiently capture global matching information across images to obtain segmentation results. Specifically, the low-dimensional index feature is used to calculate the correlation and the high-dimensional content features are combined with the correlation matrix for final segmentation. Meanwhile, we specially design a Dual-Cross Correlation Module (DCCM) which harvests the spatial and channel correlation with the adjacent pixels of another image on the cross path to enhance the representation of correlation efficiently. By utilizing a further loop operation, each feature can capture the global dependencies from all pixels of another feature. Furthermore, we fuse multi-scale correlation and features into the decoder, which is called Multi-scale Correlation Fusing Decoder (MCFD), to refine the final segmentation results. Moreover, we introduce a new dice loss function to train the whole network by averaging the dice loss value of the foreground and background. Finally, we validate our method on three co-segmentation benchmarks and the results show that our method achieves the state-of-the-art performance.**

***Index Terms*—Multi-Scale, Dual-Cross Correlation, Image Co-segmentation**

I. Introduction

Image co-segmentation is a problem of segmenting common and salient objects from a set of related images. Since this concept was firstly introduced in 2006 [1], it has attracted a lot of attention. The reasons behind its importance are two folds. On the technique aspect, the correlation between images brings valuable cues for defining the interesting objects and alleviates the ill-posed nature of segmentation. On the application aspect, image co-segmentation algorithms can be applied to various applications, such as image retrieval, Internet image processing, video tracking [2], video segmentation [3], [4], etc.

The correlation between images plus the extracted features in images provides needful cues for deciding on image co-segmentation. Most of the previous co-segmentation algorithms employed handcrafted features and correlations and embedded them into traditional computation frameworks [5]–[12]. But the algorithms based on handcrafted features and correlations suffer from their weak robustness and inexactitude. Introducing deep learning is a possible way to improve the performance of image co-segmentation through mining more sophisticated features and correlations from data. [13], [14] used deep neural networks (DNN) semantic information and constructed graphs across all correlated images as correlation information. Yuan et al. [5] considered the common class across the images as a kind of correlation. Li et al. [15], [16] represented the correlation by the patch similarity between deep features. Chen et al. [17] designed three kinds of attention pattern to get attended correlation map and forwarded it to the decoder.

Since the common objects may locate anywhere in images, we consider that image co-segmentation needs global matching between images. According to the U-Net network [18], we could find that multi-scale features are conducive to the segmentation results. Based on our observations, the existing co-segmentation methods exist the following problems: 1) Some methods could not capture the global matching information across images as the correlation computation may lead to the overflow of computational resources or the correlation obtained simply may lose important detailed information. 2) Few co-segmentation methods apply the multi-scale strategy because the correlation at a large scale also results in insufficient computing resources.

To address these problems, we propose a multi-scale Dual-Cross Correlation Network (DCNet) for image co-segmentation. Generally speaking, it is inefficient to calculate correlation directly by using the high-dimensional features extracted from Convolutional Neural Network(CNN). So we get two new Index and Content features after the CNN backbone network. One feature's channel dimension has been reduced by a factor of eight to save time and space consumption in the following operations, while the other feature reserved the channel information is used to get final segmentation results. Then, we proposed a Dual-Cross Correlation Module (DCCM) which is implemented by cross spatial and cross channel correlation with the adjacent pixels of another image on the cross path to reduce the complexity of attention process from $O(H \times W \times (H \times W))$ to $O(H \times W \times (H + W))$ comparing the attention mechanism [19]. And we repeat this module K = 2 times to reduce the localization of correlation from DCCM. Then we design a new Multi-scale Correlation Fusion Decoder

*Corresponding Author

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

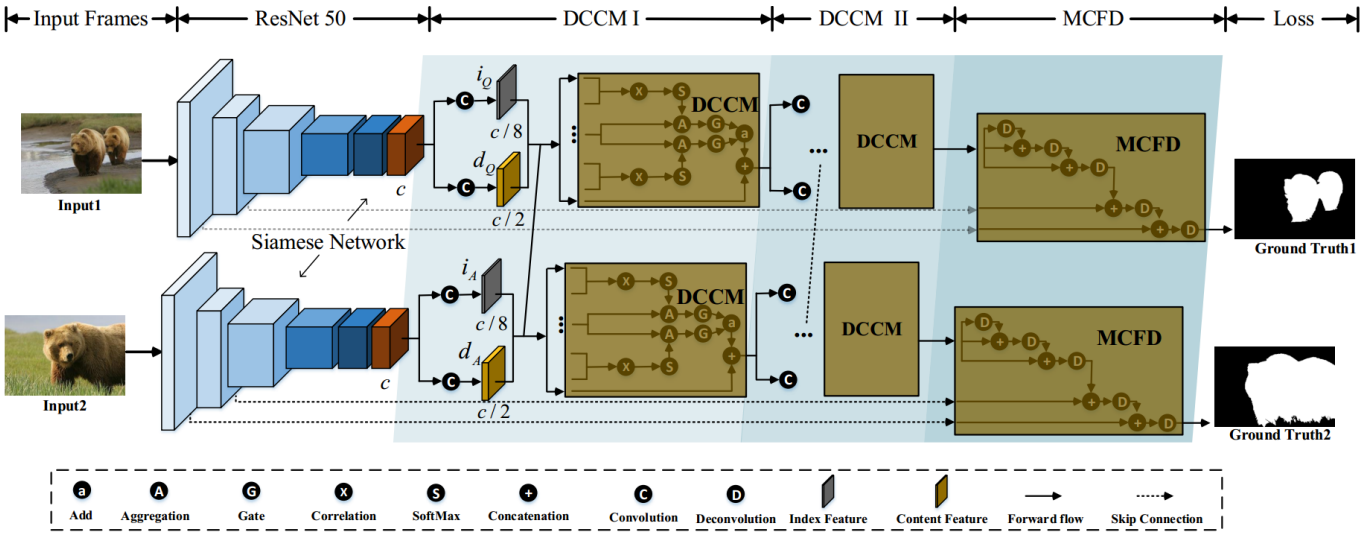

Fig. 1: Illustration of the proposed framework of Dual-Cross Correlation Network (DCNet) for image co-segmentation. Our framework includes ResNet-50 backbone, Dual-Cross Correlation Module (DCCM) and Multi-scale Correlation Fusion Decoder (MCFD) working in a Siamese manner for co-segmentation.

(MCFD) which fuses multi-scale semantic features combining correlation information produced by DCCM to refine the segmentation results via a series of pyramid deconvolution operations. Dice loss is introduced which selects the average of the foreground and background values as the total dice loss to train our framework. We evaluate the proposed co-segmentation framework on commonly used datasets iCoseg [20], Internet [21] and MSRC [22].

Our main contributions include:

- A novel multi-scale Dual-Cross Correlation Network (DCNet) is proposed to conduct high-performance image co-segmentation. To the best of our knowledge, this is the first work to propose and discuss dual correlation and multi-scale fusion in image co-segmentation.
- In DCNet, a Dual-Cross Correlation Module (DCCM) and a Multi-scale Correlation Fusion Decoder (MCFD) is proposed to enhance the representation of correlation information and fuse correlation features from different levels. A new Dice loss is also introduced to better train DCNet.
- We did extensive experiments on the popular benchmarks MSRC [22], Icoseg [20] and Internet [21], which clearly demonstrate the state-of-the-art (SOTA) performance of DCNet and effectiveness of proposed modules.

II. Related Work

A. Correlation in Traditional Co-segmentation

Traditional methods represent the correlation across images for co-segmentation based on such object elements (*e.g.* pixels, super-pixels), object regions/contours, or common object models levels. For object elements, the similarity including statistic modeling or feature space distances is usually considered to describe the intra-image and inter-image correlation among object elements [8]–[11], [13], [23]–[27] Furthermore, the correspondence between object elements in images also could be established by image matching techniques [21], [28]–[30]. For object regions/contours, the correlation can be measured by the fitting degree of object elements to object models [31] or the similarity between object models [32]–[34]. As for common object models, an optimal common object model according to all the considered images are often used to represent the common objects across the entire related images to imply the correlation between them. [6], [35]–[39].

B. Correlation in Deep Co-segmentation

For image co-segmentation based deep learning, the correlation computation methods are summarized as follows. Wang et al. [14] expressed the correlation across images by constructing an N-partite graph according to the initial segmentation results from FCN. Yuan et al. [5] used a deep network to describe the dense conditional random fields (DCRF) for computing the probability of each pixel being the foreground and obtained common objects to reflect the correlation between images. Li et al. [15] applied the Siamese network to perform image co-segmentation, where a correlation layer is employed to compute the inner product between a pixel's feature in a feature map and those of the pixels in a patch region in another feature map. [17] designed three kinds of attention pattern as the correlation between related images to get attended feature maps and forward them to the decoder. [16] utilized the similarity between deep features of different images and proposed an object proposal algorithm. Hsu et al. [40] used the normalized inner product to calculate the similarity between deep features and multiplied it with two saliencies from feature maps respectively to get the saliency guided correlation.

C. Attention Model in Computer Vision

It is well known that the attention mechanism plays a very important role in human perception [41], [42]. And the

attention mechanism has been widely used in visual task [19], [43]–[45]. Chen et al. [46] used a variety of attention maps fusing with feature maps to predict from different branches. Wang et al. [45] proposed a non-local operation to obtain the correlation between all positions in the feature map which could help to aggregate the dense contextual information. DANet [19] utilized the two types of self-attention modules to obtain contextual information which contributed to more precise segmentation results. [47] et al. proposed a memory network, in which non-local self-attention mechanism [48], [49] is extended to non-local matching to get the correlation between video frames. Lu et al. [50] proposed a co-attention module to harvest the relationship between video frames to facilitate video segmentation. Huang [51] et al. pointed out the problem of high memory occupation and time consumption in non-local attention modules [45] for the first time. And the author proposed a Criss-Cross Network (CCNet) which could capture the dependencies from all pixels and harvest the contextual information. While the Criss-Cross attention module in CCNet only considered the spatial dimension of the features and we think capturing the channel dependencies can further enhance the correlation between features.

III. Method

A. The Overall Framework

The whole deep network consists of feature extraction backbone, Dual-Cross Correlation Module (DCCM) and Multi-scale Correlation Fusing Decoder (MCFD) as shown in Figure 1. In the beginning, the inputs are two images within the same class and are fed into a two-branch Siamese backbone (ResNet-50). After that, the dense backbone features are sent to convolution layers to get content features and index features. For each branch, the features of this branch and the other branch are denoted as $\{ i^Q, d^Q \}$ (Query) and $\{ i^A, d^A \}$ (Answer) respectively. The index here can be considered as the address for retrieving specific content and indicates the importance of different locations in content features. And the content features here are high-level features with dense semantic information. Sequentially, the features $\{ i^Q, d^Q, i^A, d^A \}$ are sent to repeated DCCMs, in which the locations in the current content feature d^Q are re-weighted via a efficient dual correlation mechanism. Here in our experiments, two loops of DCCM is enough to get high performance. Finally, the re-fined features combined with the low-level features in the backbone are sent to the MCFD. In this way, low-level appearance features and high-level semantic features with correlation information are fused. The final result is obtained via a series of pyramid deconvolution operations.

B. Dual-Cross Correlation Module

Since co-segmentation aims to obtain common objects distributed in various regions from two images, we need to efficiently and globally calculate the correlation information between the two images and utilize it to guide co-segmentation. Meanwhile, we must combine the content and index features of the question image into the query image features in a reasonable way. Getting to these two points, we propose a Dual-Cross Correlation Module (DCCM), which can use the original image information and related image information to achieve promising co-segmentation performance. As shown in Figure 2, take one branch as an example, features of this branch $\{ i^Q, d^Q \}$ and the other branch $\{ i^A, d^A \}$ are inputted to DCCM of this branch. Our dual correlation includes spatial-cross correlation and channel-cross correlation, both of which bring enhancement to our method.

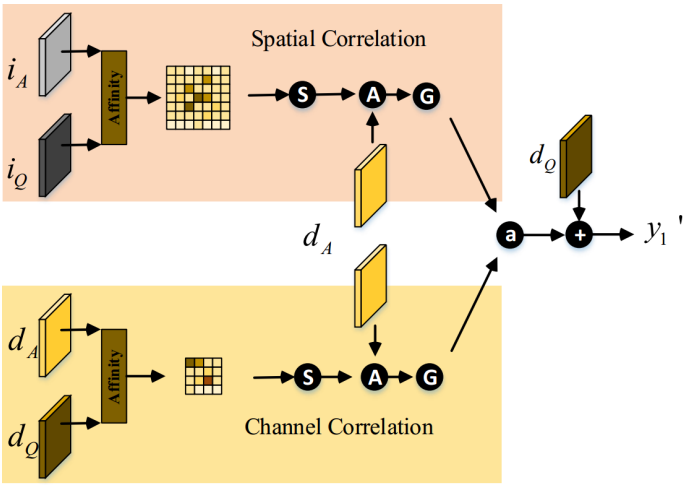

Fig. 2: The proposed Dual-Cross Correlation Module (DCCM). The spatial-cross correlation and channel-cross correlation work harmony and complement each other.

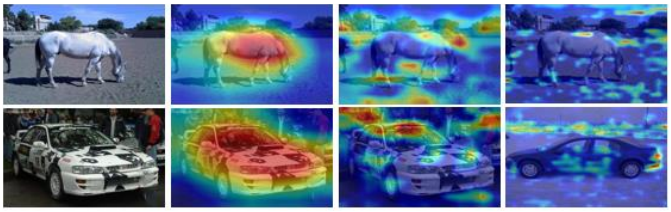

Fig. 3: Visualization of multi-scale semantic features on Internet datasets. From the first column to the last column, we show the input image, res5 features, res4 features and res3 features from ResNet-50 backbone respectively.

In both correlations, we first calculate the correlation of all pixels between the query index feature map i^Q and the answer index feature map i^A. Here we use an efficient cross-correlation mechanism based on non-local attention. As shown in Figure 1, i^Q and $i^A \in \mathbb{R}^{H\times W\times C/8}$. For each position u on spatial dimension in $i_u^A \in \mathbb{R}^{C/8}$, we can get the corresponding index feature i_u^A. Then for spatial-cross correlation, feature $S_u^S \in \mathbb{R}^{(H+W-1)\times C/8}$ can be obtained by gathering the pixels in the same column or same column with u. We can get the spatial-cross correlation on u:

$$Z_u^S = i_u^Q (S_u^S)^T, \tag{1}$$

where gathering all pixels, we can get the spatial-cross correlation $Z^S \in \mathbb{R}^{H\times W\times (H+W-1)}$. Then we do the aggregation

operation on $d_A \in \mathbb{R}^{H\times W\times C/2}$ and correlation after softmax function $A_S = softmax(Z_S)$. For each position u on d_A, we can get a cross content feature $\Omega_u \in \mathbb{R}^{(H+W-1)\times C/2}$. The long-range contextual information is collected by the aggregation operation:

$$H_u^S = \sum_{i\in|\Omega_u|} A_{i,u}\Omega_{i,u}, \tag{2}$$

where H_u^S is the corresponding spatial correlation on position u. Similarly, for the channel correlation, feature S_q^C and $S_a^C \in \mathbb{R}^{(H+W-1)\times C/2}$ can be obtained by gathering all the content feature of positions in the same channel. Then :

$$Z^C = (S_q^C)^T S_a^C, \tag{3}$$

where $Z^C \in \mathbb{R}^{C/2\times C/2}$ is the cross channel correlation. Then we do the aggregation operation on $d_A \in \mathbb{R}^{H\times W\times C/2}$ and correlation after softmax function $A_C = softmax(Z_C)$.

$$H^C = d_A A_C \tag{4}$$

There're related foreground objects in both input images. Considering that unimportant background or non-common objects between the images may negatively affect the segmentation results, we need to weight the features from different images instead of equally handling co-attention information, so we introduce a gate mechanism after obtaining H^C and H^S:

$$g_s(H^S) = \sigma(W_f \times H^S), \tag{5}$$

$$g_c(H^C) = \sigma(GAP(H^C)), \tag{6}$$

in which we denote σ as sigmoid activation function, W_f as convolutional parameters and GAP as the global average pooling operation. g_s and g_c represents the importance of different regions in H^C and H^S, so it can filter out some unnecessary information in the spatial dimension and the channel dimension, such as the background information mentioned previously or irrelevant object information. Then we can get the final output:

$$y = [H^S \odot g_s(H^S) + H^C \odot g_c(H^C), d^Q] \in \mathbb{R}^{H\times W\times C}, \tag{7}$$

where $\odot$ represents the element-wise multiplication on channel dimension and $[\cdot]$ denotes the concatenation operation.

The above process only shows the process of obtaining the correlation map by the correlation mechanism of input 1 vs. input 2. Similarly, the process of input 2 vs. input 1 uses input 2 as the query image, and the rest is the same. In the above process of obtaining y, we can see that we use combine the features of the two input images through the features of the cross-correlation map Z^S and Z^C. Compared to the full-image attention mechanism, the proposed method reduces the computing resources to a certain extent and requires less GPU memory. However, for each position in this calculation process, only the features in the same column or row are used. The non-locality of the non-local correlation mechanism will be lost, which is contradictory to the main goal of co-segmentation: to capture the common objects that appear at arbitrary positions. Thus, we need to loop the above operations to get approximately equivalent global correlations. We assume that the DCCM needs to be K times, here we take $K = 2$ as shown in Figure 2 which can get dense and rich information in two cycles.

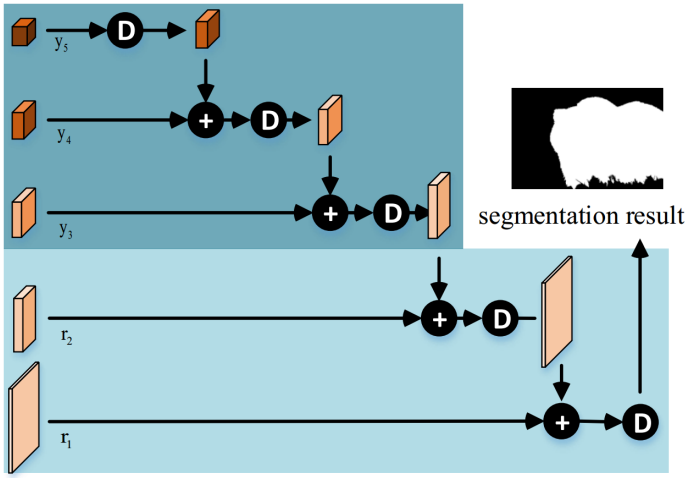

Fig. 4: Illustration of our Multi-scale Correlation Fusing Decoder (MCFD).y_3, y_4, y_5 are the outputs of the DCCM and r_1, r_2 are the outputs of the backbone network.

From the above statements, it can be seen that we can extend the cross self-attention mechanism [51] to co-segmentation through co-correlation and extend it to the combination of spatial and channel cross-correlation to better express the correlation from multi-dimension. Combined with the index feature, this mechanism can work efficiently and work complementary to the content feature. This strategy can reduce the complexity of the correlation calculation and make it easy for us to further combine it with a multi-scale strategy.

C. Multi-scale Correlation Fusing Decoder

As described above, correlation representation is sufficiently exploited in our DCCM, which can help us segment the common objects from images. Following [18], feature fusion on a single image is also the key to achieve good results in segmentation. As the visual features are shown in Figure 3, coarse high-level semantic features emphasize the abstract information of visual content, and summarize the context by large receptive field, while fine low-level visual features represent more appearance details which can better get the position of objects. Further improving the co-segmentation performance, we fuse the correlation from multiple layers to provide comprehensive representation as shown in Figure 4. In detail, we combine the multi-scale correlation feature from DCCM (y_n) with the output from the upper deconvolution layer through concatenation operation as the input of the next deconvolution layer. Specifically, we feed y_5 into the decoder as it doesn't have any upper-layer features that can be combined. At the same time, considering the complexity of the larger-scale feature correlation calculation, we do not further calculate the correlation of the larger-scale feature (res2, res1),

but combined the two features(r_1 and r_2) into the decoder directly. From y_5 features to r_1 features, the coarser feature map is upsampled by a factor of 2 using a deconvolutional layer. After the above decoder process, the final segmentation results are obtained after two deconvolutional layers. In the MCFD, except for the last deconvolution layer's kernel size is 7, the remaining deconvolution kernels size is 3. The number of channels in each deconvolution layer is set corresponding to the number of feature channels obtained by each residual block of ResNet-50.

D. Dice Loss

Image segmentation methods based on deep networks usually use the cross-entropy (CE) based learning objective. The CE measures the accuracy of binary classification of pixels. It is not directly related to the quality of segmentation. When the numbers of foreground pixels and background ones are unbalanced unless the CE is very close to zero, even seemingly nice CE could correspond to a bad segmentation. For solving the unbalanced problem, we need to seek the learning objective that can directly and more accurately reflect the quality of segmentation. Dice loss [52] is a good choice.

Let g_i^b, g_i^f be the ground truth labeling of background and foreground for the $i\text{-}th$ pixel in the image, respectively; p_i^b, p_i^f be the predicted probability of being background and foreground for the $i\text{-}th$ pixel by our network, respectively; n be the number of pixels in the image, then Dice loss measured on background and foreground category is:

$$DL^b = 1 - \frac{2\sum_{i=1}^{n} g_i^b \cdot p_i^b + \varepsilon}{\sum_{i=1}^{n} g_i^b + \sum_{i=1}^{n} p_i^b + \varepsilon}, \tag{8}$$

$$DL^f = 1 - \frac{2\sum_{i=1}^{n} g_i^f \cdot p_i^f + \varepsilon}{\sum_{i=1}^{n} g_i^f + \sum_{i=1}^{n} p_i^f + \varepsilon} \tag{9}$$

respectively, where ε is a small value for preventing zero denominators. The total dice loss is the mean of the two losses on the background and foreground.

IV. Experiments

A. Experimental Setup

Datasets In image co-segmentation community, iCoseg [20], Internet [21], and MSRC [22] are widely used as evaluation dataset. However, the number of annotated examples in these datasets is limited and is not enough to train deep networks. The bigger PSACAL VOC 2010 and 2012 dataset are recently used as the training sets in image co-segmentation methods involving deep networks [5], [14], [15], [53]. We take MSRC and VOC 2012 [54] as training sets. MSRC is composed of 591 images of 21 object groups. The ground-truth is roughly labeled, which does not align exactly with the object boundaries. VOC 2012 include 11540 images with ground-truth detection boxes and 2913 images with segmentation masks. Only 2913 images with segmentation masks can be considered in our problem. Note that not all of the examples in these two datasets can be used. In MSRC, some images include only stuff and some are used as the test subset images. In VOC 2012, the interested objects in some images have great changes in appearance and are cluttered in many other objects, so that the meaningful correlation between them is too hard to be found. The remaining 1743 images in VOC 2012 and 430 images in MSRC are used to construct our training set. From the training images, we sampled 13200 pairs of images containing common objects to train our co-segmentation network.

TABLE I: The performance comparisons on MSRC-subset.

Methods	[57]	[12]	[10]	[58]	[15]	[17]	Ours
P	90.2	92.2	92.0	84.0	92.4	95.3	**96.0**
J	0.71	0.75	0.77	0.67	0.80	0.78	**0.85**

Evaluation Metrics Two commonly used metrics for segmentation evaluation are used: Precision and Jaccard index. Precision (denoted by P) is the percentage of correctly classified pixels in both the background and foreground. Jaccard index (denoted by J) is the overlapping rate of foreground between segmentation result and ground truth mask.

Implementation Details We make the input images to 448×448 sizes considering the limited computing resource and producing 28×28 feature maps by ResNet-50 backbone. The Deep features Extraction ResNet-50 are initialized with weights trained on the Imagenet dataset [55]. We use Adam [56] optimizer with learning rate 1e-5 for optimization and the weight decay of $5e$-5. The training process takes about 30 hours using a single NVIDIA TITAN XP GPU. When evaluating our method on the test dataset, we randomly select a related image to compose an image pair with the test image.

B. Comparison to the State-of-the-Arts

We compare our DCNet on the MSRC dataset, the Internet dataset, and the iCoseg dataset. The performance from previous counterparts based on deep networks and the previous best one from traditional methods are both included for comparison.

Table I and Table II show our competitive results on the MSRC subsets and the Internet dataset which we call it the Seen Class dataset. We can see that DCNet gets the state of the art performance when segmenting Seen class Objects. On the MSRC dataset, we use the subsets which are not included in our training sets has 7 classes and 10 images in each class. From Table I, we find that DCNet improves the performance in both precision and Jaccard index by increase rates 6.3% in J and the rate 0.7% in P comparing the second-best method [15], [17]. Following the compared methods, we evaluate DCNet on the Internet widely-used subset, in which each class has 100 images. From the results in Table II, we see that DCNet outperforms the second-best results [15], [17] by the large increase rates 9.5% in J and the rate 2.5% in P, respectively. Figure 5 shows some examples of co-segmentation results by DCNet for each category on the Internet dataset. We can see

TABLE II: The performance comparisons on Internet.

Method	Car		Horse		Airplane		Average	
	P	J	P	J	P	J	P	J
[27]	88.0	0.71	88.3	0.60	90.5	0.61	88.9	0.64
[5]	90.4	0.72	90.2	0.65	92.6	0.66	91.1	0.68
[53]	88.7	0.68	89.3	0.58	92.3	0.60	90.1	0.62
[15]	94.0	0.83	91.4	0.65	94.6	0.64	93.3	0.71
[17]	-	0.80	-	0.71	-	0.71	-	0.74
[16]	93.0	0.82	89.7	0.61	94.2	0.67	92.3	0.70
Ours	**96.9**	**0.92**	**93.3**	**0.72**	**96.5**	**0.80**	**95.6**	**0.81**

that DCNet can accurately segment the common objects under various appearances, poses, and backgrounds clutter.

Fig. 5: Examples of co-segmentation results on internet dataset. From the first row to the last row, the classes are Airplane, Car and Horse respectively.

Following the previous work [15], [17], we use the iCoseg subset to evaluate DCNet. This dataset contains 8 classes with different image numbers. As the class in iCoseg dataset is different from the training dataset, so many methods use the dataset to test the generalizability of their models. So we consider the iCoseg subset as Unseen Objects. According to the Jaccard Index results from Table III, DCNet outperforms on all class results and the average result which means it could also adapt to the various unseen class data. Especially, our result on the "Cheetah" class improve 16.7% comparing the second-best result and improve 5.8% on the average value. Furthermore, only 13200 training samples are used in this work comparing [17] 160k training samples. This shows DCNet can achieve better results with fewer data and less training time. Figure 6 shows some examples of co-segmentation results on iCoseg dataset by DCNet. We can see that the method accurately segment the interested objects with an accurate edge, which can adapt to the changes in the size, the pose, and the number of interested objects.

DCNet can get better results in all three benchmarks, the main reasons are as follows: 1) The efficient DCCM proposed in our paper can well represent the correlation between the related images and help us get the fine segmentation result. 2) The combination of multi-scale correlation into the decoder part in DCNet plays an important role in fining the edge of the segmentation result. 3) Dice loss can obtain the segmentation results by optimizing directly on the segmentation target, which is better than cross-entropy.

C. Ablation Studies

To better verify DCNet, we conduct extensive ablation experiments on the Internet dataset and iCoseg dataset with different modules for our method.

TABLE III: The comparisons of Jaccard index on iCoseg-subset (Unseen Class).

Class	[10]	[27]	[15]	[17]	Ours
Bear2	0.70	0.68	0.88	0.88	**0.91**
Brownbear	0.92	0.73	0.92	0.92	**0.93**
Cheetah	0.67	0.78	0.69	0.71	**0.91**
Elephant	0.67	0.80	0.85	0.84	**0.90**
Helicopter	0.82	0.80	0.79	0.77	**0.82**
Hotballoon	0.88	0.80	0.92	0.94	**0.94**
Panda1	0.70	0.72	0.83	0.92	**0.94**
Panda2	0.55	0.61	0.87	0.90	**0.93**
Average	0.78	0.74	0.84	0.86	**0.91**

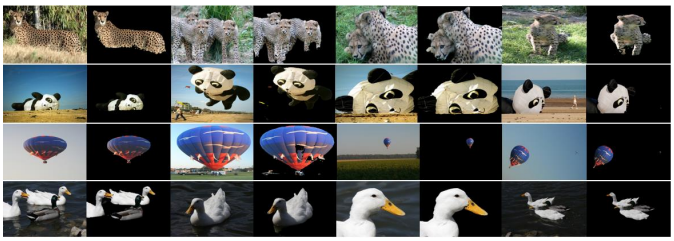

Fig. 6: Examples of co-segmentation results on iCoseg dataset. From the first row to the last row, the classes are Cheetah, Kitepanda, Hotballoon, and Goose respectively.

Baseline In our baseline, the overall network mainly includes three parts: feature extraction network (ResNet-50), correlation read module (CRM) and decoder as shown in Figure 1. We first selected the non-local attention following [45] which is to implement the global matching between images in the correlation read module. We also use the index and content features to input into the CRM and get the retrieval weight by calculating the non-local correlation of the two index features. The weighted content features of the answer image retrieved by the correlation matrix are concated with content features of the query image. We combine the single-scale (the last scale of ResNet-50 feature map) correlation feature output by CRM with the feature of ResNet-50 each residual block to obtain the segmentation result.

The effect of Multi-scale Fusion Decoder Based on the baseline, we fuse the multi-scale correlation feature to the decoder part, and the way of feature fusion is shown in Figure 4. We selected the features of the three scales of ResNet-50 and used CRM to obtain relevant features for segmentation. In Table IV, MCFD(+) shows that compared with the segmentation results obtained at the baseline of a single scale, the multi-scale fusion decoder improves the performance by the increase rates 3% in J and the rate 0.5% in P. The result can prove that combining high-level abstract feature correlation with low-level visual feature correlation can help us refine our segmentation results as high-level abstract semantic features contain more semantic information, while low-level visual features contain more detailed appearance information.

The effect of Dual-Cross Correlation Module We replaced the CRM module in the multi-scale baseline (described

TABLE IV: Ablation study of our method on the Internet dataset.

Method	P	J
Baseline	94.1	0.77
MCFD(+)	94.6	0.79
SCCM(+)	94.9	0.80
DCCM(+)	95.6	0.81

TABLE V: Ablation study of our DCCM on the iCoseg-subset.

Method	P	J
SCCM(+)	97.4	0.89
DCCM(+)	97.6	0.91

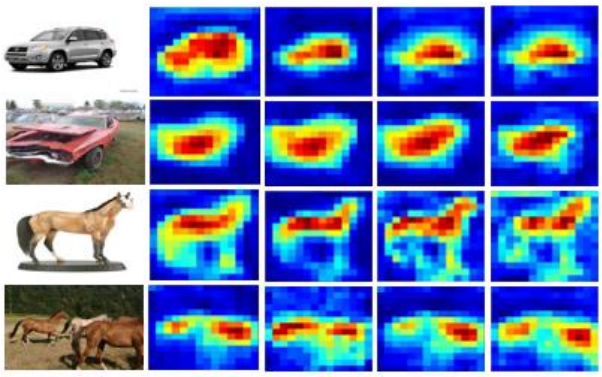

Fig. 7: Visualization results of DCCM modules on the Internet datasets. For each row, we show input image, four spatial-cross correlation maps corresponding to the input image and we input the top and bottom two rows of the same class image into our network at the same time.

as MCFD(+) in the table) with our proposed DCCM module (also called DCNet) and conducted ablation experiments. It can be seen from SCCM(+) result in Table IV that the spatial cross-criss module (SCCM) [51] which we modify it for DCNet to fit our CRM feature fusion strategy can improve the performance of Internet dataset by the increase rates 1.3% in J and the rate 0.3% in P comparing with MCFD(+). After combining the spatial-cross correlation with the channel-cross correlation as our Dual-Cross Correlation Module (DCCM) called DCCM(+) in Table IV, we find that our performance is improved by the increased rates 1.3% in J and the rate 0.7% in P comparing to SCCM(+) which only use spatial correlation on the Internet dataset. From the result of the iCoseg-subset in Table V, our DCCM outperform the SCCM(+) [51] by the increase rate 2.2% in J and 0.2% in P. We can conclude that the channel-cross correlation added into DCCM can complement spatial-cross correlation by enhancing the representation of image correlation information and our DCCM really work in image co-segmentation task.

Visualization of Correlation Map To better understand our DCCM and verify its role, we visualize the spatial-cross correlation features of the DCCM output in Figure 7 as the channel-cross correlation features are not easily visualized in the spatial dimension. We can see from the visual map from fusion features, the correlation feature has a high response value at the location of the object. For example, the horses of the input image pair are both highlighted by the correlation visual map more than one horse appears in one image.

V. Conclusion

This paper proposes a novel multi-scale Dual-Cross Correlation Network (DCNet) for performing image co-segmentation, which is constructed by introducing Dual-Cross Correlation Module (DCCM) and Multi-scale Correlation Fusing Decoder (MCFD) to help us obtain the final segmentation results better. To obtain better learning results, Dice loss is introduced to further improve the performance. The ablation experiments demonstrate that the DCCM can effectively enhance the expression of related information compared with spatial-cross correlation, as well as the MCFD can further refine the segmentation results based on the single-scale decoder. Compared with the previous method including traditional and deep learning methods, the proposed DCNet achieved the state-of-the-art performance on three representative image co-segmentation datasets. For further improving the performance of our approach in future work, we plan to extend the acquisition of the dual-cross correlation to a set of related images, rather than limited to a pair of images.

References

[1] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, "Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2006, pp. 993–1000.

[2] X. Lu, C. Ma, B. Ni, X. Yang, I. Reid, and M.-H. Yang, "Deep regression tracking with shrinkage loss," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 353–369.

[3] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3623–3632.

[4] W. Le, H. Gang, R. Sukthankar, J. Xue, and N. Zheng, "Video object discovery and co-segmentation with extremely weak supervision," in *Proceedings of the European Conference on Computer Vision*, 2014.

[5] Z.-H. Yuan, T. Lu, and Y. Wu, "Deep-dense conditional random fields for object co-segmentation." in *IJCAI*, 2017, pp. 3371–3377.

[6] C. Lee, W.-D. Jang, J.-Y. Sim, and C.-S. Kim, "Multiple random walkers and their application to image cosegmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3837–3845.

[7] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 169–176.

[8] L. Liu, K. Li, and X. Liao, "Image co-segmentation by co-diffusion," *Circuits, Systems, and Signal Processing*, vol. 36, no. 11, pp. 4423–4440, 2017.

[9] L. Li, Z. Liu, and J. Zhang, "Unsupervised image co-segmentation via guidance of simple images," *Neurocomputing*, vol. 275, pp. 1650–1661, 2018.

[10] A. Faktor and M. Irani, "Co-segmentation by composition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1297–1304.

[11] E. Kim, H. Li, and X. Huang, "A hierarchical image clustering cosegmentation framework," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 686–693.

[12] J. Ma, S. Li, H. Qin, and A. Hao, "Unsupervised multi-class co-segmentation via joint-cut over l_1-manifold hyper-graph of discriminative image regions," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1216–1230, 2017.

[13] R. Quan, J. Han, D. Zhang, and F. Nie, "Object co-segmentation via graph optimized-flexible manifold ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 687–695.

[14] C. Wang, H. Zhang, L. Yang, X. Cao, and H. Xiong, "Multiple semantic matching on augmented n-partite graph for object co-segmentation," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5825–5839, 2017.

[15] W. Li, O. H. Jafari, and C. Rother, "Deep object co-segmentation," in *Asian Conference on Computer Vision*, 2018.

[16] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Co-attention cnns for unsupervised object co-segmentation." in *IJCAI*, 2018, pp. 748–756.

[17] H. Chen, Y. Huang, and H. Nakayama, "Semantic aware attention based deep object co-segmentation," in *Asian Conference on Computer Vision*, 2018, pp. 435–450.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015, pp. 234–241.

[19] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[20] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "icoseg: Interactive co-segmentation with intelligent scribble guidance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3169–3176.

[21] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in internet images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1939–1946.

[22] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Proceedings of the European Conference on Computer Vision*, 2006, pp. 1–15.

[23] T. Taniai, S. N. Sinha, and Y. Sato, "Joint recovery of dense correspondence and cosegmentation in two images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4246–4255.

[24] F. Meng, H. Li, G. Liu, and K. N. Ngan, "Object co-segmentation based on shortest path algorithm and saliency model," *IEEE Transactions on Multimedia*, vol. 14, no. 5, pp. 1429–1441, 2012.

[25] X. Liang, L. Zhu, and D.-S. Huang, "Multi-task ranking svm for image cosegmentation," *Neurocomputing*, vol. 247, pp. 126–136, 2017.

[26] Z. Tao, H. Liu, H. Fu, and Y. Fu, "Image cosegmentation via saliency-guided constrained clustering with cosine similarity." in *AAAI*, 2017, pp. 4285–4291.

[27] K. R. Jerripothula, J. Cai, and J. Yuan, "Image co-segmentation via saliency co-fusion," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1896–1909, 2016.

[28] K. R. Jerripothula, J. Cai, J. Lu, and J. Yuan, "Object co-skeletonization with co-segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3881–3889.

[29] J. Cech, J. Matas, and M. Perdoch, "Efficient sequential correspondence selection by cosegmentation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 32, no. 9, pp. 1568–1581, 2010.

[30] J. C. Rubio, J. Serrat, A. López, and N. Paragios, "Unsupervised co-segmentation through region matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 749–756.

[31] F. Meng, H. Li, G. Liu, and K. N. Ngan, "Image cosegmentation by incorporating color reward strategy and active contour model," *IEEE transactions on cybernetics*, vol. 43, no. 2, pp. 725–737, 2013.

[32] Z. Wang and R. Liu, "Semi-supervised learning for large scale image cosegmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 393–400.

[33] L. Mukherjee, V. Singh, and J. Peng, "Scale invariant cosegmentation for image groups," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011, p. 1881.

[34] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 2129–2136.

[35] Y. Li, J. Liu, Z. Li, H. Lu, and S. Ma, "Object co-segmentation via salient and common regions discovery," *Neurocomputing*, vol. 172, pp. 225–234, 2016.

[36] W. Wang and J. Shen, "Higher-order image co-segmentation," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1011–1021, 2016.

[37] H. Zhu, J. Lu, J. Cai, J. Zheng, and N. M. Thalmann, "Multiple foreground recognition and cosegmentation: An object-oriented crf model with robust higher-order potentials," in *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, 2014, pp. 485–492.

[38] J. Sun and J. Ponce, "Learning discriminative part detectors for image classification and cosegmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3400–3407.

[39] G. Kim and E. P. Xing, "On multiple foreground cosegmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 837–844.

[40] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Deepco^3: Deep instance co-segmentation by co-peak search and co-saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[41] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 11, pp. 1254–1259, 1998.

[42] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature reviews neuroscience*, vol. 3, no. 3, p. 201, 2002.

[43] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[44] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[45] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[46] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3640–3649.

[47] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," *arXiv preprint arXiv:1904.00607*, 2019.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[49] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," *arXiv preprint arXiv:1601.06733*, 2016.

[50] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3623–3632.

[51] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 603–612.

[52] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571.

[53] J. Han, R. Quan, D. Zhang, and F. Nie, "Robust object co-segmentation using background prior," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1639–1651, 2018.

[54] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[57] S. Vicente, C. Rother, and V. Kolmogorov, "Object cosegmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 2217–2224.

[58] P. Mukherjee, B. Lall, and S. Lattupally, "Object cosegmentation using deep siamese network," *arXiv preprint arXiv:1803.02555*, 2018.