Image and Vision Computing 101 (2020) 103973

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

A novel co-attention computation block for deep learning based image co-segmentation

Xiaopeng Gong, Xiabi Liu *, Yushuo Li, Huiyu Li

Beijing Lab of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China

ARTICLE INFO

Article history:
Received 18 December 2019
Received in revised form 23 June 2020
Accepted 25 June 2020
Available online 07 July 2020

Keywords:
Visual co-attention
Image co-segmentation
Deep learning
Correlation calculation
Average pooling

ABSTRACT

The correlation between images is crucial for solving the image co-segmentation problem that is segmenting common and salient objects from a set of related images. This paper proposes a novel co-attention computation block to compute the visual correlation between images for improving the co-segmentation performance. Here 'co-attention' means that we obtain the co-attention features in encoded features of an image to guide the attention in another image. To this purpose, we firstly introduce top-k average pooling to compute the channel co-attention descriptor. Then we explore the correlation between features in different spatial positions to get the spatial co-attention descriptor. Finally, these two types of co-attention descriptors are multiplied to generate a fused one. We obtain such a fused co-attention descriptor for each image and use it to produce the co-attention augmented feature map for the following processing in the applications. We embed the proposed co-attention block into a U-shaped Siamese network for fulfilling the image co-segmentation. It is proven to be able to improve the performance effectively in the experiments. To our best knowledge, it leads to the currently best results on Internet dataset and iCoseg dataset.

1. 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 attentions and many co-segmentation algorithms have been proposed. The reasons behind its importance are two folds. On the technique aspect, the correlation between images brings valuable cues for defining the interested objects and alleviates the ill-pose nature of segmentation. On the application aspect, image co-segmentation algorithms can be applied to and play crucial roles in various applications, such as Internet image mining, image retrieval, video tracking, video segmentation, and etc.

The visual/semantic object features in images plus the correlation between images provide needful cues for image cosegmentation. Most of the previous co-segmentation algorithms explored handcrafted object features and correlations and conducted the computation on the level of object elements (such as pixels, superpixels, or over-segmented regions) [2–7], object regions/contours [8–12], or common object models [9,13–15]. Although much progress has been made, the algorithms based on handcrafted object features and correlations still suffer from their unrobustness and inexactitude. Recently, deep learning was introduced to improve the performance of image co-segmentation through mining more sophisticated object features and correlations from data [16–20].

The correlation between images brings the main advantage of alleviating the ill-pose nature of co-segmentation, compared with other segmentation problems without such information. In this paper, we propose a co-attention computation block to compute the correlation between images for deep learning based image co-segmentation. Here 'co-attention' means that the attention in an image is determined by the related information in another image. To our best knowledge, this is the first introduction of such co-attention schema for exploring the relevance among multiple images in deep visual analysis. Our co-attention computation block includes channel co-attention module and spatial co-attention module, called CC-A and SC-A for short, respectively. In CC-A, we introduce top-k average pooling to measure the semantic response in each channel. It cooperates with a gating function to generate a channel co-attention descriptor for an image. In SC-A, the feature map of each image is convolved to one-dimensional firstly; then each two one-dimensional features from two images are multiplied to measure the spatial correlation between the two images; finally, the resultant spatial correlation map is convolved to one-dimensional map that passes through a gating function to produce the spatial co-attention descriptor. CC-A and SC-A reflect 'what' and 'where' the images' features should be emphasized or suppressed, respectively. We multiply them together to obtain a fused one that is used to process the feature maps of the images to get the co-attention from the view of both channel and space. We embed the proposed

* Corresponding author.
E-mail address: liuxiabi@bit.edu.cn (X. Liu).

https://doi.org/10.1016/j.imavis.2020.103973
0262-8856/

co-attention computation block into a U-shaped Siamese network for performing image co-segmentation.

Our main contributions are summarized as follows:

(1) Co-attention computation block: the idea of determining the attention in an image according to the related information in another image is firstly introduced and implemented in deep network structure. Accordingly, a novel co-attention computation block is presented to obtain the correlation map between images;
(2) Top-k average pooling: we extend the global average pooling widely used in attention community to top-k average pooling, which is more flexible and more accurate for grasping the semantics hidden in the channel responses;
(3) The application to image co-segmentation: we apply the proposed co-attention computation block to image co-segmentation and achieve the currently best results on commonly used test datasets.

2. Related work

2.1. Correlation calculation in traditional image co-segmentation methods

In accordance with the strategies of traditional methods of image co-segmentation introduced in the last section, the correlation between images can be measured in the level of object elements, object regions/contours, or common object models. For object elements, we usually establish an objective function, mostly often energy function, to describe the intra-image and inter-image relationship among object elements. By optimizing such objective function, we determine the labeling of each element in each image. The similarity between object elements is mainly considered to reflect the correlation between images, which is computed through statistic modeling or feature distances [2,4–6,10,11,17,21–27]. Furthermore, we can also establish the correspondence between object elements in images for reflecting the correlation, by using image matching techniques [15,28–30]. For object regions/contours, a model is used to represent object regions or contours. We try to find out the optimal models which are fitted to the images for completing the segmentation. The correlation can be measured by the fitting degree of object elements to object models [8], or the similarity between object models [1,12,31–34], or the correspondence between objects in different images [35]. For common object models, they are used to model the common objects across the images. The difference between common object models and object region/contour models is that a common object model keeps same for the entire related images while object region/contour models are specific to single images and vary across the images. We try to find out an optimal common object model according to all the considered images. Such model implies the correlation between the images [3,14,36–41].

2.2. Correlation calculation in deep learning based image co-segmentation methods

For deep learning based image co-segmentation, the correlation computation methods are summarized as follows. Yuan et al. [16] obtained common objects to reflect the correlation between images. They used a deep network to describe the dense conditional random fields (DCRF) for the common object in the images. Such a deep network is used to compute the probability of each pixel being the foreground, which was used in the second DCRF procedure to decide the final labels of pixels. Wang et al. [17] measured the correlation between images as the similarities between the initial segmentation results from fully convolutional network (FCN). They used FCN to obtain the initial segmentation result on each single image. Then the co-occurrence of candidate regions is computed by using N-partite graph method. Finally, the GrabCut is applied on the co-occurrence map to obtain the segmentation results. Han et al. [18] computed the similarities in visual and semantic features between superpixels to reflect the correlation. They constructed two graphs according to low-level visual features and high semantic features. The high semantic features were computed by using a convolutional neural network (CNN). On these two graphs, two probability maps corresponding to the segmentation results were computed. Li et al. [19] applied Siamese network to co-segmentation, which is composed of three parts: Siamese encoder for using CNN network to extract features from the two images, the mutual correlation for matching two image features, and Siamese decoder for obtaining segmentation results by using the deconvolutional computation. Hsu et al. [20] used the normalized inner product to compute the similarity between two features and multiplied it with two saliency computed in two feature maps respectively to get the saliency guided correlation. Correspondingly, they decomposed the image co-segmentation task into two sub-tasks, co-peak search and instance mask segmentation. In the former sub-task, the joint co-peak and co-saliency map are detected. In the latter sub-task, the high-quality instance proposals are retrieved for accomplishing instance co-segmentation.

2.3. Co-attention mechanism

It is well known that the attention plays an important role in human perception [42–44]. Recently, the co-attention between texts and images has attracted interests in deep textual-visual analysis. You et al. [45] selected semantic concepts in Neural Image Caption (NIC) and computed the confidences for the feature map by using attribute classifiers. Jia et al. [46] exploited the correlation between images and their captions as the global semantic information to guide the long short-term memory (LSTM) network generating sentences. Yang et al. [47] proposed a stacked attention network that queries the image multiple times to infer the answer progressively. Nam et al. [48] proposed a dual attention network, which combines the visual and textual attention via multiple reasoning steps to adaptively integrate local features with their global dependencies.

The concept of 'co-attention' in these previous works is different from ours, which denotes the mutual influence between texts and images to form text-guided visual attention or image-guided textual attention. While the 'co-attention' in this paper means the mutual influence of two images, where the attentive features obtained in the encoded features of an image is used to determine the attention in another image.

Furthermore, Fu et al. [49] also proposed to use channel attention module and spatial attention module in their method for capturing rich contextual dependencies (DANet). However, they considered 'self-attention' mechanism instead of 'co-attention'. In their problem, only one image is involved. Thus, DANet captured the spatial dependencies and channel dependencies by multiplying a feature map with its own transpose. Differently, the 'co-attention' is this paper involves multiple features from different images.

3. The proposed method

In this section, our co-attention computation block and its integration with U-shaped Siamese network for image co-segmentation are presented. We firstly describe the channel co-attention, the spatial co-attention, and the fusion of them. Then we introduce the whole architecture of our image co-segmentation network with the proposed co-attention block.

3.1. Channel co-attention (CC-A) module

The features in each channel of a feature map in deep networks are obtained by a specific convolution kernel that is learned from the training data. Thus different channels in the feature map can reflect different

semantics. Also, CC-A can reflect 'what' the images' features should be emphasized. We observe the image aroused responses in each channel and find out that there is a positive connection between the responses in the channels and the specific semantic: the images from the same class arouse the largest response in a certain channel and the small responses in other channels; and different classes correspond to different channels. Fig. 1 illustrates an example of such situation, where we randomly select the three channels in the final feature map of ResNet-50 network to observe their responses to different classes of images. We can see that each of three channels reflects the semantic content of a certain class, respectively.

This is the reason why we can use global average pooling [50], which takes the average of each channel over the feature map and generates a vector, to reflect the semantics hidden in the channel response. We find that the global average pooling can be extended to top-k average pooling for getting more accurate results, which means larger values in each channel, instead of all the values in it, are averaged over the feature map. Here, top-k means top k percent of largest values in each channel. Let A be the considered feature map, whose width, height, and channel number are W, H, and C, A_c be the c-th channel of A, A_c^i be the i-th maximum value in A_c, $0 < k \le 1$ be the percentage of top values that we are interested in, then the top-k average pooling for A_c is computed by using

$$TopK(A_c) = \frac{1}{k \times H \times W} \sum_{i=1}^{k \times H \times W} A_c^i. \tag{1}$$

Such computation is more flexible than global average pooling. Actually, the global average pooling is a special case of top-k average pooling with $k = 1$. Theoretically speaking, the small responses in each channel should correspond to the background if the segmentation is good, so selecting top-k maximum values and ignoring the other values for each channel is helpful to reduce the influence of the background for getting more robust segmentation results. So, by setting up an appropriate k value, we can make the computation focus on the main semantic part of the images and reduce the disturbance of the unrelated information such as backgrounds. In this paper, the value of k is set up through careful experiments.

The results from top-k average pooling are inputted into a gating function to get the attentive outputs of each channel, called channel co-attention descriptor. This descriptor expresses the global distribution of channels' responses, in which each value reflects the probability that the image has certain semantic content.

Given a pair of images, we compute the channel co-attention descriptor described above for each image. Then the channel co-attention descriptor of one image is used to multiply with the feature map of another image to obtain the correlation map based on channel co-attention. The overall procedure of this computation is summarized in the network block shown in Fig. 2 and can be described formally as

$$Y_A = \sigma(TopK(A)), \tag{2}$$

$$Y_B = \sigma(TopK(B)), \tag{3}$$

$$A' = A \otimes Y_B, \tag{4}$$

$$B' = B \otimes Y_A, \tag{5}$$

where $\sigma(\cdot)$ is the gating function (softmax is used in this paper); $\otimes$ denotes the element-wise multiplication using broadcast mechanism; A and B are the encoded features of two images; Y_A and Y_B are channel co-attention descriptors; A' and B' are resultant correlation maps from channel co-attention.

3.2. Spatial co-attention (SC-A) module

Channel attention shows us what features we should concern in the images. Besides this, we need to determine where to focus on, i.e., to find out the attentive parts in the space of images. Fig. 3 shows an example of spatial attention in an image where the attentive parts are shown in heat map.

Our network block of computing spatial co-attention is illustrated in Fig. 4. Let A and B be a pair of feature maps of two images. These two feature maps A and B represent the high-level semantic content of the input images. When the two images contain objects that belong to a common class, they should contain similar features at the locations of the shared objects. Therefore, we employ the "product" operation to

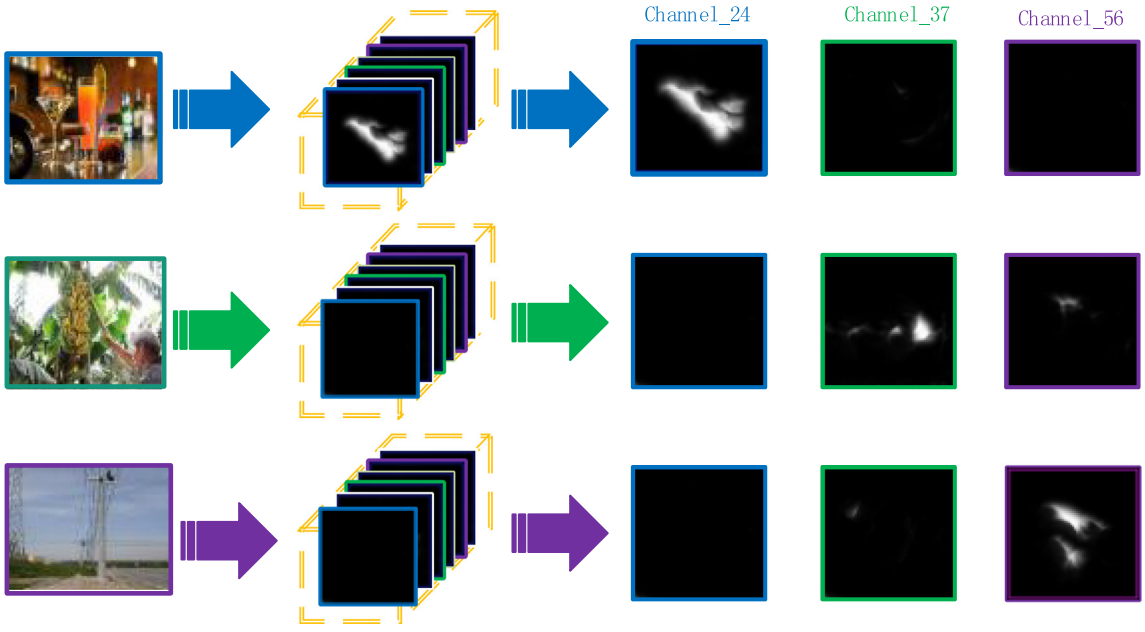

Fig. 1. The illustration of the connection between semantics and channels of the final feature map in well-trained deep networks (taking Resnet-50 as an example), where the intensity of pixels in gray level images indicates the channels' response values aroused by color images at the left. The same colors of the borders of images, the arrows and the texts indicate the relationship among the images, their categories, and the representation of the categories in the feature map. For example, the blue corresponds to the category 'Cup'.

Fig. 2. The processing diagram of CC-A, where Y_A and Y_B are the channel co-attention descriptors.

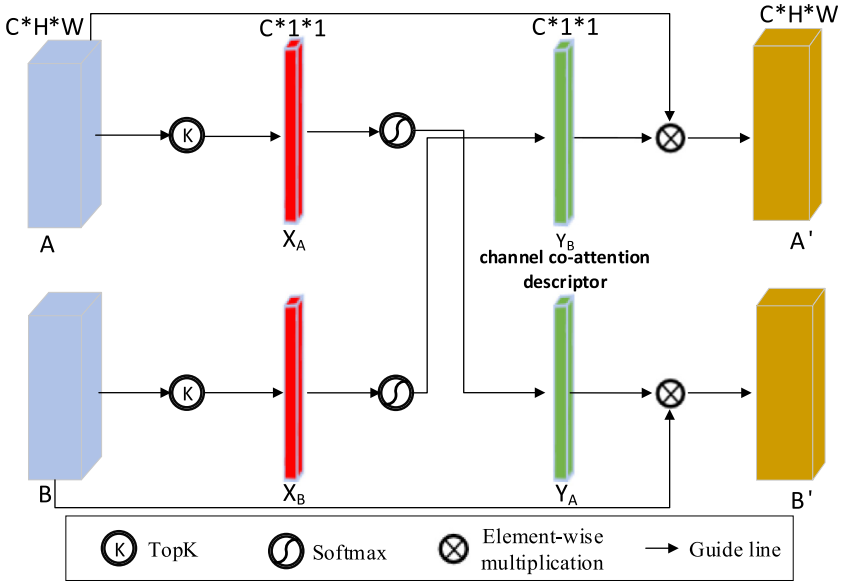

Fig. 3. The illustration of spatial attention, where the main attentive parts corresponding to the concept of 'surfing' are shown in heat map.

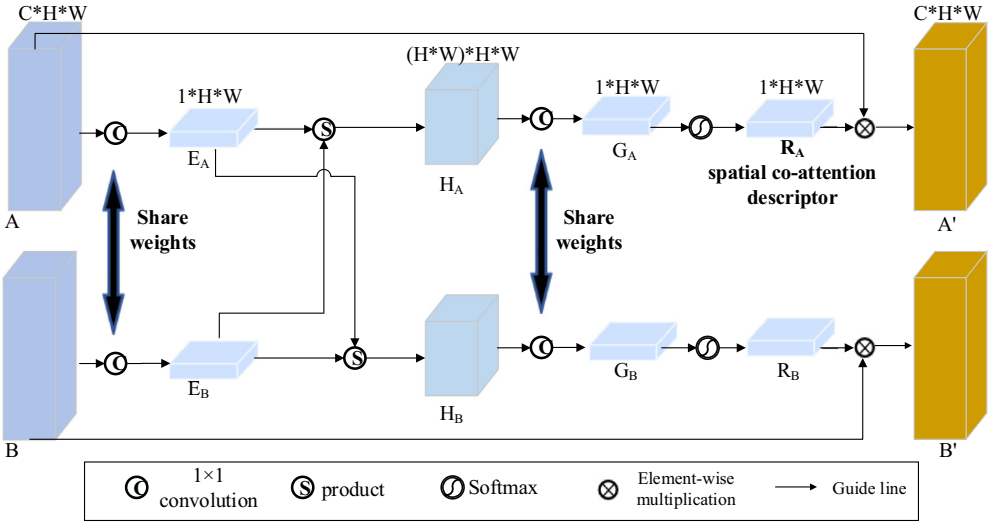

Fig. 4. The processing diagram of SC-A. where R_A and R_B are the spatial co-attention descriptors.

perform a pixel-wise comparison between two feature maps for computing the correlation between each pair of locations on the feature maps. First, in order to reduce the computational complexity, we apply a 1×1 kernel to convolve the feature maps into one-dimensional ones; let the results from A and B be denoted as E_A and E_B, respectively. Second, each two features in E_A and E_B are multiplied

to reflect the similarity between them. In order to obtain the spatial information, each point in E_A should be relevant with all points in E_B, so that this product can be implemented by taking each point in E_A as a kernel to convolve E_B. Let the resultant feature map be denoted as $H_A \in R^{(W\times H)\times W\times H}$. The same processing is conducted for E_B to generate a feature map H_B. Third, H_A and H_B are convolved into a one-dimensional feature map again, respectively. The one-dimensional feature maps pass through a gating function to get our spatial co-attention descriptors, which actually capture the overall context dependency between each pixel in one image and all the pixels in another image. Finally, multiplying spatial co-attention descriptors with the corresponding original feature maps (A or B) yields the correlation map from spatial co-attention.

The computation procedure described above can be summarized formally as

$$C_{AB} = S(Conv(W_1, A), Conv(W_1, B)), \tag{6}$$

$$C_{BA} = S(Conv(W_1, B), Conv(W_1, A)), \tag{7}$$

$$R_A = \sigma(Conv(W_2, C_{AB})), \tag{8}$$

$$R_B = \sigma(Conv(W_2, C_{BA})), \tag{9}$$

$$A' = A \otimes R_A, \tag{10}$$

$$B' = B \otimes R_B, \tag{11}$$

where the meaning of $\sigma(\cdot)$ and $\otimes$ are same as those in Eqs. (2)–(5); W_1 and W_2 are the weights of the kernels in the first and the second layer of convolution in our SC-A, respectively; $Conv(a, b)$ denotes the convolutional result of taking a as one convolution kernel and b as one feature map. $S(\cdot, \cdot)$ denotes the multiplication of two one-dimensional feature maps; R_A and R_B are the resultant spatial co-attention descriptors; A' and B' are resultant correlation maps from spatial co-attention.

3.3. Fused co-attention (FC-A) module

CC-A represents the semantic distribution in an image, each value in which reflects the intensity of a specific semantics in the image. Therefore, the possibility of having the same semantics in two images can be measured by multiplying CC-A from an image with the feature map of another image. SC-A addresses the problem where the common features between images are, each value in which reflects the similarity between two features at corresponding positions of two images. So, SC-A considers the positions of common features in two images but ignores the semantics; while CC-A considers the common semantics between two images but neglects their space information. Intuitively, we can strength CC-A by multiplying it with SC-A to decide what and where the common semantics are in two images. According to the above discussions, CC-A and SC-A can be applied to measure the relevance between two images for improving the segmentation results, respectively, and can be fused to obtain more better results.

We fuse them in the following way. First, we multiply the channel co-attention descriptor with the spatial co-attention descriptor to get a fused one. Then, this fused co-attention descriptor is multiplied with the feature maps to get the fused co-attention. This combination processing is illustrated in Fig. 5.

3.4. Extension to multi-images

The above co-attention operations are explained from the view of two images. It can be extended to the cases of more than two images. Let $I = \{I_1, \ldots, I_N\}$ be the set of input images. We pair each image in I with other Q images ($Q \le N - 1$) in it. Then, we use the method described in the last subsection to compute fused co-attention descriptors for each pair; let the results be denoted as $\hat{A} = \{A_n^q : 1 \le n \le N, 1 \le q \le Q\}$, and F_n denotes the mean of all the fused co-attention descriptors. We can integrate all the fused co-attention descriptors for each image through averaging, i.e., we have

$$F_n(x, y) = \frac{1}{Q}\sum_{q=1}^{Q} A_n^q(x, y) \tag{12}$$

for each image. The median can make our approach more robust. So this fused co-attention descriptor can be used to broadcast the attention to groups with outliers in the corresponding images.

3.5. Image co-segmentation architecture and its learning

We apply the proposed co-attention block to image co-segmentation. It is embedded into a U-shaped Siamese network. The overall architecture of resultant network is shown in Fig. 6, which is composed of three parts. The first part is Siamese encoders that are a pair of two feature encoder networks, each of which extracts semantic features from an image, respectively. The two encoders share weights with each other. ResNet [51] introduced residual blocks to make deeper networks without reducing accuracy and behaved well in image classification and object detection. Most recently, it also demonstrated good results in video segmentation [52,53]. Thus we adopt ResNet-50 network [51] to construct our Siamese encoder. The second part is our proposed co-attention block, through which the co-attention based correlation maps are calculated from the two semantic feature maps. This makes our network be different from other Siamese networks for image co-segmentation, such as [19]. The third part is Siamese decoders, which are constructed by symmetrically reversing each of two encoders and injecting each scale of semantic features into the reversed pathway. In this part, the co-attention features are concatenated with semantic features to proceed decoding. In this way, the semantics of objects and the co-attention between objects are jointly utilized to detect common and salient objects across the images.

Our network is trained in an end-to-end way. Given pairs of training images labeled with segmentations, all the weights in the network, including those of our co-attention block, are optimized simultaneously with Stochastic Gradient Descent (SGD) to minimize the Dice loss [54] that is same as Jaccard index explained in the experiments.

4. Experiments

We conduct two groups of experiments to evaluate the effectiveness of our proposed co-attention block. In the first group, the performance of our image co-segmentation network with the proposed co-attention block is tested and compared with other six state-of-the-art techniques including Jerripothula et al. [7], Ma et al. [6], Faktor and Irani [26], Han et al. [18], Yuan et al. [16] and Li et al. [19]. These methods reported the previous best performance among deep methods or traditional non-deep methods, on the experimental data sets used in this paper. Please refer to Section 2 for introduction to these methods. In the second group of experiments, the design of our co-attention block is justified through ablation study.

4.1. Experimental setup

4.1.1. Datasets

We use Pascal VOC 2012[1] [18] and MSRC[2] [55] datasets to train our image co-segmentation network, and then use Internet [29] and iCoseg [3] as the test sets. These four data sets are widely used in the community of image co-segmentation. We further test our network on the

[1] Pattern analysis statistical modeling and computational learning, visual object classes, 2012.
[2] Microsoft Research Cambridge.

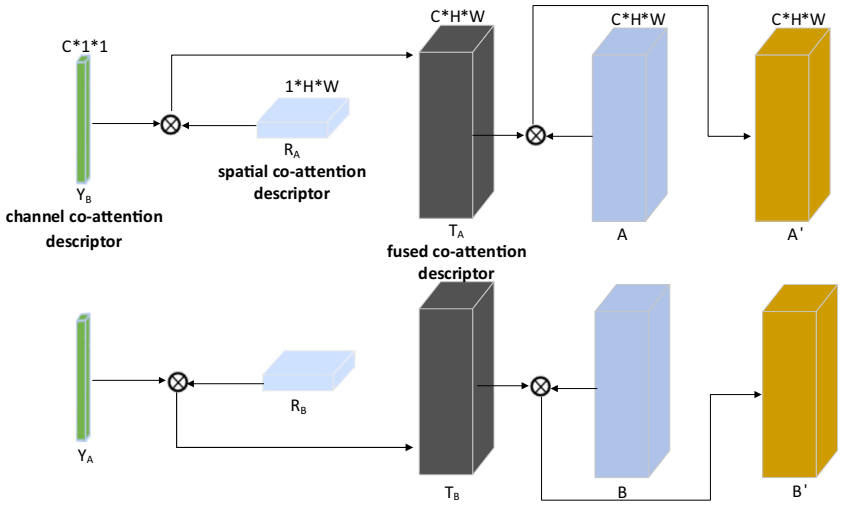

Fig. 5. The processing diagram of FC-A, where T_A and T_B are the fused co-attention descriptors by multiplying the channel co-attention descriptor with the spatial co-attention descriptor.

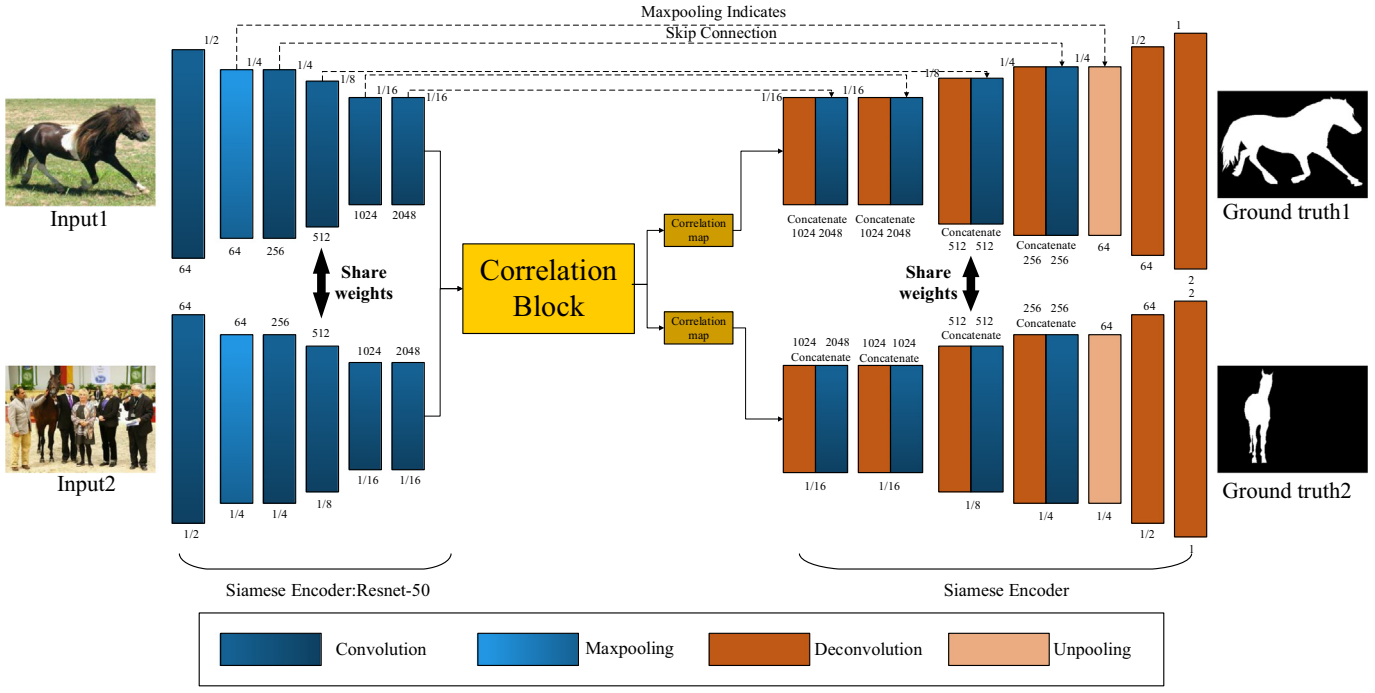

Fig. 6. The architecture of our U-shaped Siamese network with the proposed co-attention block for image co-segmentation.

Cityscapes [56] dataset, which is commonly used for semantic segmentation in recent years, but has not been seen used for co-segmentation.

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 includes 11,540 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 without obvious foreground, such as only sky or grassland. 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 ambiguous. We exclude them from consideration. The remained 1743 images in VOC 2012 and 507 images in MSRC are used to construct our training set. From the training images, we sampled 41,329 pairs of images containing common objects to train our proposed image co-segmentation network.

iCoseg dataset contains 643 images divided into 38 object groups. Each group contains 17 images on average. The pixel-wise hand-annotated ground truth is offered. The backgrounds in each group are consistent natural scenes. Besides entire iCoseg dataset, many previous works use various versions of its subsets. Internet consists of 3 classes (airplane, car, and house) of thousands of downloaded Internet images. Following the compared methods, we evaluate our approach on its widely used subset, in which each class has 100 images.

Cityscapes dataset has 5000 finely annotated images, involving 19 semantic classes. These images are divided into 2975, 500, and 1525 images for training, validation, and testing, respectively. The annotations of test images are withheld for benchmark. It has

Table 1
The performance comparisons on Internet.

Method	Car		Horse		Airplane		Average	
	Precision	Jaccard	Precision	Jaccard	Precision	Jaccard	Precision	Jaccard
Jerripothula et al. [7]	88.0	0.71	88.3	0.60	90.5	0.61	88.9	0.64
Han et al. [18]	88.7	0.68	89.3	0.58	92.3	0.60	90.1	0.62
Yuan et al. [16]	90.4	0.72	90.2	**0.65**	92.6	0.66	91.1	0.68
Li et al. [19]	94.0	0.83	91.4	**0.65**	94.6	0.64	93.3	0.71
Ours	**94.7**	**0.87**	**93.3**	**0.65**	**95.5**	**0.76**	**94.5**	**0.76**

The bold numbers indicates the best results among all methods.

another 20 k coarse annotations for training, which are not used in our experiments.

4.1.2. Evaluation metrics

We use two commonly used metrics for evaluating the effects of image co-segmentation: Precision and Jaccard index.

Precision is the percentage of correctly classified pixels in both background and foreground, which can be defined as

$$Precision = \frac{|Segmentation \cap Ground\ truth|}{|Segmentation|}. \quad (13)$$

Jaccard index (denoted by Jaccard in the following descriptions) is the overlapping rate of foreground between the segmentation result and the ground truth mask, which can be defined as

$$Jaccard = \frac{Segmentation \cap Ground\ truth}{Segmentation \cup Ground\ truth}. \quad (14)$$

4.1.3. Parameter setting

We conduct the experiments on a computer with GTX 1080Ti GPU and implement the image co-segmentation network with PyTorch. In the experiments, the batch size for training is set to be 16, the learning rate is initialized to 0.02 and is divided by 10 at the 8-th and the 11-th epochs, and the weight decay and the momentum parameters are set to be 10, 4 and 0.9, respectively. The optimization procedure ends after 40 epochs. Because of limited computing resource, all images are resized to the resolution of 448 × 448 in advance. The co-segmentation results are resized back to the original image resolution for performance evaluation.

4.2. Comparisons on the internet dataset

The resultant performances on the Internet dataset from our method (denoted by OURS) as well as the compared methods are listed in Table 1. Each value in Table 1 is the mean in ten times for each method. The reason why Faktor and Irani [26] and Ma et al. [6] methods are missed here is that they didn't report the results on the Internet dataset.

Table 2
The performance comparisons on iCoseg-entire.

Method	Precision	Jaccard
Faktor and Irani [26]	92	0.73
Ma et al. [6]	–	0.79
Han et al. [18]	94.4	0.78
Yuan et al. [16]	94.4	0.82
Ours	**95.3**	**0.83**

Table 3
The performance comparisons on iCoseg-subset1.

Method	Precision	Jaccard
Faktor and Irani [26]	94.4	0.79
Han et al. [18]	94.8	0.85
Yuan et al. [16]	96.0	**0.86**
Ours	**96.2**	**0.86**

The bold numbers indicates the best results among all methods.

From the results shown in Table 1, we can conclude that our method outperforms the currently best methods, not only deep network based ones but also traditional ones. Such improvement occurs on all three classes and in both precision and Jaccard index. Compared with the previous best one (from Li et al. [19]), the increase rates in precision and average Jaccard index brought by our method are 1.3% and 7.0%, respectively. In addition, the minimum, maximum and standard deviation of ten precisions are 93.9, 95.1 and 0.48132, respectively, and those of ten Jaccard indexes are 0.755, 0.764 and 0.00474, respectively, which shows that our results are hardly affected by random factors and that our network is stable.

Fig. 7 shows some examples of segmented images from our method. It can be seen that our network can accurately segment the common objects under various appearances, poses, and backgrounds. It can also deal with the pairs of unrelated images. As shown in the last column of each case, for those images that are unrelated to the corresponding categories, almost all the pixels are segmented as background.

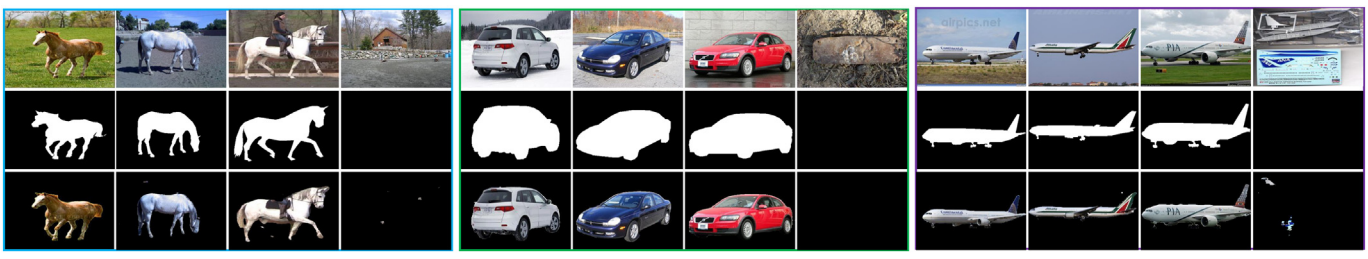

Fig. 7. Co-segmentation results of the cases corresponding to horse, car, and airplane from Internet dataset, where the top row shows the images that are paired with each other as the input of our network; the middle row shows the ground-truths; and the bottom row shows our segmented results.

Table 4
The comparisons of Jaccard index on iCoseg-subset2.

Class	Faktor and Irani [26]	Jerripothula et al. [7]	Li et al. [19]	Ours
Bear2	0.70	0.68	0.88	0.89
Brownbear	0.92	0.73	0.92	0.94
Cheetah	0.67	0.78	0.69	0.78
Elephant	0.67	0.80	0.85	0.88
Helicopter	0.82	0.80	0.79	0.81
Hotballoon	0.88	0.80	0.92	0.94
Panda1	0.70	0.72	0.83	0.86
Panda2	0.55	0.61	0.87	0.88
Average	0.78	0.74	0.84	**0.87**

The bold numbers indicates the best results among all methods.

4.3. Comparisons on the iCoseg dataset

We evaluate the proposed approach on not only the entire iCoseg, but also on its two subsets: iCoseg-subset1 with 16 classes and iCoseg-subset2 with 8 classes. The resultant performances are listed in Tables 2–4, respectively. Each value in Tables 2–4 is the mean in ten times for each method. As done in Internet case, we compare our performance with previous best ones from the methods based on deep learning and previous best ones from traditional methods. From the results shown in these three Tables, we can conclude that our method stably outperforms state-of-the-art counterparts. It brings the best precision on both iCoseg-entire and iCoseg-subset1. As for Jaccard index, our approach leads to the best one in entire iCoseg and iCoseg-subset2, and shares the first place in the iCoseg-subset1. It should be noted that only 2250 annotated images are used in this work. This amount is much less than 8498 used in Yuan et al. [16]. It is hopeful to further improve the performance of our approach by increasing the amount of training data. In addition, the minimum, maximum and standard deviation of ten precisions on iCoseg-entire are 94.8, 95.7 and 0.32147, respectively, and those of ten Jaccard indexes are 0.826, 0.834 and 0.00309, respectively, which again illustrates that our network model is very stable.

Fig. 8 shows some examples of good co-segmentation results on iCoseg dataset by using our method. We can see that our method accurately segment the interested objects, which can adapt to the changes on the size, the pose, and the number of interested objects.

Fig. 9 shows some failed results. As shown in Fig. 9, for the class 'Airshowsplanes', the smokes ejected by the airplanes are easy confused with the airplanes themselves, so that these smokes are often falsely decided as foreground; for the class 'WomanSoccers', since our network doesn't aim at instance-level segmentation, we fail to identify adjoining objects from the same class.

4.4. Test on the cityscapes dataset

In the test, we use finely annotated training set and validation set to train the network and test the final model, respectively. Since the co-segmentation task is targeted to segment a group of common images, we randomly combine two images which have the common foreground. Finally, 5950, 1500 pairs of images are generated in training set, validation set, respectively.

We also conduct the same experiments in ten times. In ten experiments, our method attains 91.3, 90.8, 92.0 and 0.25463 for the mean, minimum, maximum and standard deviation of precision on validation set, respectively, and 0.84, 0.835, 0.844 and 0.00451 for those of Jaccard index, respectively. To our knowledge, this is the first use of Cityscapes in testing image co-segmentation, so there is no comparison between our results and others.

We randomly choose some examples from the validation set and visualize the segmented results in Fig. 10. It can be seen that although the objects are small, our network can segment them well.

Fig. 8. Examples of correct co-segmentation results on iCoseg dataset, where the 1st and 4th rows show origins; the 2nd and 5th rows show ground-truths; the 3rd and 6th rows show results.

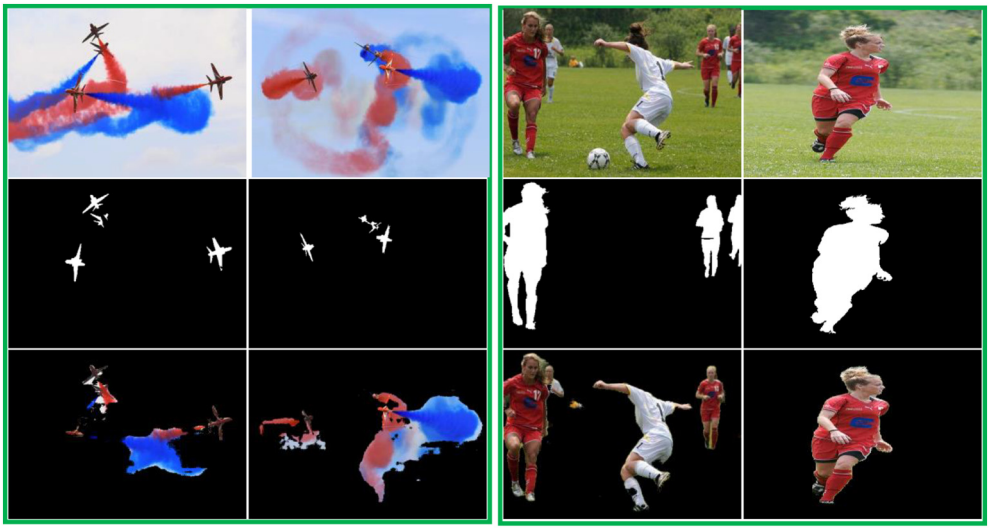

Fig. 9. Examples of un-correct co-segmentation results on iCoseg dataset, where the 1st row shows origins; the 2nd shows ground-truths; the 3rd shows results.

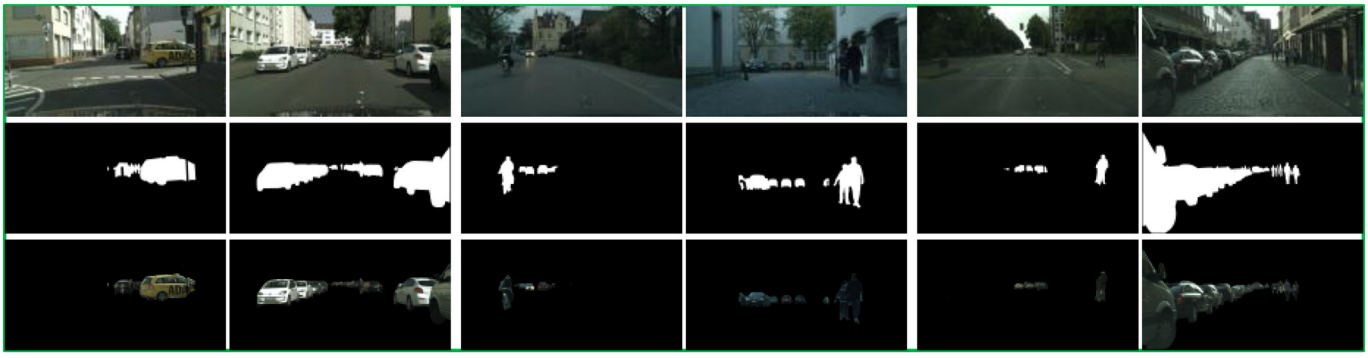

Fig. 10. Examples of our co-segmentation results on Cityscapes dataset, where the 1st row shows origins; the 2nd row shows ground-truths; the 3rd row shows results.

4.5. Ablation study

In order to justify the design of our co-attention computation block, we make the following changes to our image co-segmentation network and compare the performance of these modified versions: 1) Baseline: only considering the inner product between feature maps to construct the correlation layer; 2) CC-A(+): replacing the baseline correlation layer with our channel co-attention module; 3) SC-A(+):replacing the baseline correlation layer with our spatial co-attention module; 4) FC-A(+): replacing the baseline correlation layer with our fused channel and spatial co-attention module.

We repeat the experiments ten times on the Internet in these ablation studies and record the mean values. The training sets are still VOC 2012 and MSRC. The performance of these modified versions of our network is shown in Table 5, which demonstrates that: 1) compared with the Baseline, both channel co-attention and spatial co-attention can bring useful correlation information for improving the performance; 2) The combination of channel co-attention and spatial co-attention further improves the performance. This confirms the reasonability of the design of our co-attention computation block.

The inner product used in our Baseline network is actually a main method of measuring the correlation between two images in existing deep learning based co-segmentation methods, such as in Li et al. [19]. From the view of theory, the inner product reflects the similarity between features in two images, thus it is similar with our spatial co-attention descriptor (SC-A) for grasping the spatial aspect of common objects but ignoring their semantic aspect. Furthermore, it is not based

Table 5
The comparisons of ablated and completed network on Internet.

Method	Precision	Jaccard
Baseline	93.3	0.71
CC-A(+)	94.1	0.74
SC-A(+)	93.8	0.72
FC-A(+)	**94.5**	**0.76**

The bold numbers indicates the best results among all methods.

Table 6
The performance of top-k average pooling on Internet using FC-A.

k	Precision	Jaccard	k	Precision	Jaccard
0.1	93.4	0.75	0.6	93.9	0.76
0.2	93.5	0.75	0.7	94.1	0.76
0.3	93.4	0.74	**0.8**	**94.5**	**0.76**
0.4	93.5	0.75	0.9	94.3	0.76
0.5	93.7	0.76	1.0	93.9	0.75

The bold numbers indicates the best results among all methods.

on the attention mechanism like our SC-A. So the effectiveness of the inner product is not only worse than our fused one but also those from our SC-A and CC-A.

Furthermore, in order to test the effectiveness of top-k average pooling and determine the best k, we conduct the experiments with different k values from 0.1 to 1.0 under the FC-A version of our network. The mean resultant performance on ten times is shown in Table 6. As shown there, we can see that all in $k = 0.6 \sim 0.9$ outperforms the global average pooling (i. e., $k = 1.0$) and $k = 0.8$ leads to the best result. Actually, k is set to be 0.8 in all the experiments reported above.

5. Conclusions

In this paper, we have introduced the co-attention idea of determining the attention in an image through the related information in another image and proposed a co-attention computation block to improve the performance of deep learning based image co-segmentation. We design the channel co-attention module, the spatial co-attention module, and the combination strategy of them. The ablation study demonstrates that both channel co-attention and spatial co-attention are useful, and the fused co-attention is more better. The experiments of image co-segmentation on commonly used datasets confirm the effectiveness of the proposed co-attention computation block. Compared with the state-of-the-art methods based on deep learning or traditional techniques, the proposed approach achieves the currently best performance on Internet dataset and iCoseg dataset. It proves its ability of extracting related information and excluding unrelated information between images. In the future, we will try to determine the approbierte k value for top-k average pooling automatically and use more training data to further improve the performance of our image co-segmentation network.

Author statement

Xiaopeng gong: Conceptualization, Methodology, Software, Validation, Visualization, Writing - Original Draft, Resources, Writing - Review & Editing.

Xiabi Liu: Project administration, Writing - Review & Editing.

Yushuo Li: Investigation, Resources, Data Curation, Supervision.

Huiyu Li: Investigation, Resources, Data Curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China [grant number 81171407] and the Beijing Municipal Science and Technology Project [grant number Z181100001918002].

References

[1] Carsten Rother, Tom Minka, Andrew Blake, Vladimir Kolmogorov, Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs, 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1, IEEE 2006, pp. 993–1000.

[2] Gunhee Kim, Eric P. Xing, Li Fei-Fei, Takeo Kanade, Distributed cosegmentation via submodular optimization on anisotropic diffusion, 2011 International Conference on Computer Vision, IEEE 2011, pp. 169–176.

[3] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, Tsuhan Chen, icoseg: Interactive co-segmentation with intelligent scribble guidance, 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE 2010, pp. 3169–3176.

[4] Armand Joulin, Francis Bach, Jean Ponce, Discriminative clustering for image co-segmentation, 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE 2010, pp. 1943–1950.

[5] Edward Kim, Hongsheng Li, Xiaolei Huang, A hierarchical image clustering cosegmentation framework, 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE 2012, pp. 686–693.

[6] Jizhou Ma, Shuai Li, Hong Qin, Aimin Hao, Unsupervised multi-class co-segmentation via joint-cut over l_{1}-manifold hyper-graph of discriminative image regions, IEEE Trans. Image Process. 26 (3) (2016) 1216–1230.

[7] Koteswar Rao Jerripothula, Jianfei Cai, Junsong Yuan, Image co-segmentation via saliency co-fusion, IEEE Transactions on Multimedia 18 (9) (2016) 1896–1909.

[8] Fanman Meng, Hongliang Li, Guanghui Liu, King Ngi Ngan, Image cosegmentation by incorporating color reward strategy and active contour model, IEEE Transactions on Cybernetics 43 (2) (2013) 725–737.

[9] Jifeng Dai, Ying Nian Wu, Jie Zhou, Song-Chun Zhu, Cosegmentation and cosketch by unsupervised learning, Proceedings of the IEEE International Conference on Computer Vision 2013, pp. 1305–1312.

[10] Fanman Meng, Jianfei Cai, Hongliang Li, Cosegmentation of multiple image groups, Comput. Vis. Image Underst. 146 (2016) 67–76.

[11] Fanman Meng, Hongliang Li, Guanghui Liu, King Ngi Ngan, Object co-segmentation based on shortest path algorithm and saliency model, IEEE Transactions on Multimedia 14 (5) (2012) 1429–1441.

[12] Lopamudra Mukherjee, Vikas Singh, Charles R. Dyer, Half-integrality based algorithms for cosegmentation of images, 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE 2009, pp. 2028–2035.

[13] Wenguan Wang, Jianbing Shen, Higher-order image co-segmentation, IEEE Transactions on Multimedia 18 (6) (2016) 1011–1021.

[14] Hongyuan Zhu, Jiangbo Lu, Jianfei Cai, Jianming Zheng, Nadia M. Thalmann, Multiple foreground recognition and cosegmentation: an object-oriented crf model with robust higher-order potentials, IEEE Winter Conference on Applications of Computer Vision, IEEE 2014, pp. 485–492.

[15] Jose C. Rubio, Joan Serrat, Antonio López, Nikos Paragios, Unsupervised co-segmentation through region matching, 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE 2012, pp. 749–756.

[16] Ze-Huan Yuan, Tong Lu, Yirui Wu, Deep-dense conditional random fields for object co-segmentation, IJCAI (2017) 3371–3377.

[17] Chuan Wang, Hua Zhang, Yang Liang, Xiaochun Cao, Hongkai Xiong, Multiple semantic matching on augmented n-partite graph for object co-segmentation, IEEE Trans. Image Process. 26 (12) (2017) 5825–5839.

[18] Junwei Han, Rong Quan, Dingwen Zhang, Feiping Nie, Robust object co-segmentation using background prior, IEEE Trans. Image Process. 27 (4) (2017) 1639–1651.

[19] Weihao Li, Omid Hosseini Jafari, Carsten Rother, Deep object co-segmentation, Asian Conference on Computer Vision, Springer 2018, pp. 638–653.

[20] Kuang-Jui Hsu, Yen-Yu Lin, Yung-Yu Chuang, Deepco3: deep instance co-segmentation by co-peak search and co-saliency detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, pp. 8846–8855.

[21] Rong Quan, Junwei Han, Dingwen Zhang, Feiping Nie, Object co-segmentation via graph optimized-flexible manifold ranking, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 687–695.

[22] Sara Vicente, Carsten Rother, Vladimir Kolmogorov, Object cosegmentation, CVPR 2011, IEEE 2011, pp. 2217–2224.

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

[24] Xianpeng Liang, Zhu Lin, De-Shuang Huang, Multi-task ranking svm for image cosegmentation, Neurocomputing 247 (2017) 126–136.

[25] Fu Huazhu, Dong Xu, Stephen Lin, Jiang Liu, Object-based rgbd image co-segmentation with mutex constraint, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, pp. 4428–4436.

[26] Alon Faktor and Michal Irani. Co-segmentation by composition. In 2013 IEEE International Conference on Computer Vision.

[27] Zhiqiang Tao, Hongfu Liu, Fu Huazhu, Fu. Yun, Image cosegmentation via saliency-guided constrained clustering with cosine similarity, Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[28] Koteswar Rao Jerripothula, Jianfei Cai, Jiangbo Lu, Junsong Yuan, Object co-skeletonization with co-segmentation, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE 2017, pp. 3881–3889.

[29] Michael Rubinstein, Armand Joulin, Johannes Kopf, Ce Liu, Unsupervised joint object discovery and segmentation in internet images, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2013, pp. 1939–1946.

[30] Jan Cech, Jiri Matas, Michal Perdoch, Efficient sequential correspondence selection by cosegmentation, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1568–1581.

[31] Dorit S. Hochbaum, Vikas Singh, An efficient algorithm for co-segmentation, 2009 IEEE 12th International Conference on Computer Vision, IEEE 2009, pp. 269–276.

[32] Kai-Yueh Chang, Tyng-Luh Liu, Shang-Hong Lai, From co-saliency to co-segmentation: an efficient and fully unsupervised energy minimization model, CVPR 2011, IEEE 2011, pp. 2129–2136.

[33] Lopamudra Mukherjee, Vikas Singh, Jiming Peng, Scale invariant cosegmentation for image groups, CVPR 2011, IEEE 2011, pp. 1881–1888.

[34] Zhengxiang Wang, Rujie Liu, Semi-supervised learning for large scale image cosegmentation, Proceedings of the IEEE International Conference on Computer Vision 2013, pp. 393–400.

[35] Fan Wang, Qixing Huang, Leonidas J. Guibas, Image co-segmentation via consistent functional maps, Proceedings of the IEEE International Conference on Computer Vision 2013, pp. 849–856.

[36] Chulwoo Lee, Won-Dong Jang, Jae-Young Sim, Chang-Su Kim, Multiple random walkers and their application to image cosegmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, pp. 3837–3845.

[37] Yong Li, Jing Liu, Zechao Li, Hanqing Lu, Songde Ma, Object co-segmentation via salient and common regions discovery, Neurocomputing 172 (2016) 225–234.
[38] Zhao Liu, Jianke Zhu, Bu Jiajun, Chun Chen, Object cosegmentation by nonrigid mapping, Neurocomputing 135 (2014) 107–116.
[39] Jian Sun, Jean Ponce, Learning discriminative part detectors for image classification and cosegmentation, Proceedings of the IEEE International Conference on Computer Vision 2013, pp. 3400–3407.
[40] Yuning Chai, Esa Rahtu, Victor Lempitsky, Luc Van Gool, Andrew Zisserman, Tricos: a tri-level class-discriminative co-segmentation method for image classification, European Conference on Computer Vision, Springer 2012, pp. 794–807.
[41] Gunhee Kim, Eric P. Xing, On multiple foreground cosegmentation, 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE 2012, pp. 837–844.
[42] Laurent Itti, Christof Koch, Ernst Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis & Machine Intelligence, 11, 1998, pp. 1254–1259.
[43] Ronald A. Rensink, The dynamic representation of scenes, Vis. Cogn. 7 (1–3) (2000) 17–42.
[44] Maurizio Corbetta, Gordon L. Shulman, Control of goal-directed and stimulus-driven attention in the brain, Nature Reviews Neuroscience 3 (3) (2002) 201.
[45] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, Jiebo Luo, Image captioning with semantic attention, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 4651–4659.
[46] Jia Xu, Efstratios Gavves, Basura Fernando, Tinne Tuytelaars, Guiding the long-short term memory model for image caption generation, Proceedings of the IEEE International Conference on Computer Vision 2015, pp. 2407–2415.
[47] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola, Stacked attention networks for image question answering, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 21–29.
[48] Hyeonseob Nam, Jung-Woo Ha, Jeonghee Kim, Dual attention networks for multimodal reasoning and matching, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 299–307.
[49] Fu Jun, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, Hanqing Lu, Dual Attention Network for Scene Segmentation, 2019 3146–3154.
[50] Jie Hu, Li Shen, Gang Sun, Squeeze-and-excitation networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 7132–7141.
[51] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 770–778.
[52] Yuan-Ting Hu, Jia-Bin Huang, Alexander G. Schwing, Videomatch: matching based video object segmentation, Proceedings of the European Conference on Computer Vision (ECCV) 2018, pp. 54–70.
[53] Oh. Seoung Wug, Joon-Young Lee, Kalyan Sunkavalli, Seon Joo Kim, Fast video object segmentation by reference-guided mask propagation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 7376–7385.
[54] Fausto Milletari, Nassir Navab, Seyed-Ahmad Ahmadi, V-net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation, arXiv preprint, 2016arXiv:1606.04797.
[55] Jamie Shotton, John Winn, Carsten Rother, Antonio Criminisi, Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation, European Conference on Computer Vision, Springer 2006, pp. 1–15.
[56] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, Bernt Schiele, The Cityscapes Dataset for Semantic Urban Scene Understanding, 2016.