Learn Image Object Co-segmentation with Multi-scale Feature Fusion

Lina Li
Shanghai University
Shanghai, China
University of Technology, Sydney
Sydney, Australia
Lina.Li-1@student.uts.edu.cn

Zhi Liu
Shanghai University
Shanghai, China
liuzhisjtu@163.com

Jian Zhang
University of Technology, Sydney
Sydney, Australia
Jian.Zhang@uts.edu.au

Xiaofei Zhou
Hangzhou Dianzi University
Hangzhou, China
zxforchid@outlook.com

***Abstract*—Image object co-segmentation aims to segment common objects in a group of images. This paper proposes a novel neural network, which extracts multi-scale convolutional features at multiple layers via a modified VGG network and fuses them both within and across images as the intra-image and the inter-image features. Then these two kinds of features are further fused at each scale as the multi-scale co-features of common objects, and finally the multi-scale co-features are summed up and upsampled to obtain the co-segmentation results. To simplify the network and reduce the rapidly rising resource cost along with the inputs, the reduced input size, less downsampling and dilation convolution are adopted in the proposed model. Experimental results on the public dataset demonstrate that the proposed model achieves a comparable performance to the state-of-the-art co-segmentation methods while the computation cost has been effectively reduced.**

***Index Terms*—Image co-segmentation, object co-segmentation, multi-scale, multi-layer, dilated convolution.**

I. Introduction

Image object co-segmentation aims to segment the common objects in a group of images that contain the same or similar objects, and researchers have paid sustained attention to it and the related research areas such as image co-saliency detection [1]–[5] and image object co-localization [6]. Generally, the common object segmentation needs both intra-image object probability computation and inter-image object probability computation, as the common object regions not only get high object probabilities but also share high inter-image similarities.

There have been quite a few methods for image co-segmentation [7]–[14]. Most of the methods are unsupervised, and the common objects are discovered based on some low-level handcrafted features such as luminance, colors or textures. The similarities among the common objects are computed based on the pixels or regions. In [7], the objects are segmented from the background by clustering the image pixels into two clusters that can be maximally distinguished. In [1]–[3], [9], [10], region similarities are computed among all the images of the group, while in [8], regions are compared only within the selected images which share the most similarities with the target image. In [12], regions are compared only within the selected simple images. In [11], guided by quality measurement, saliency maps are fused from different images using the dense correspondence.

This work was supported by the National Natural Science Foundation of China under Grants 61771301 and 61901145.

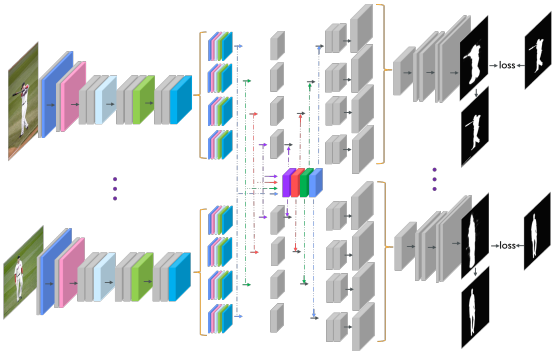

Fig. 1. An overview of the proposed model. It can be viewed as three parts. The input images are first passed through the modified VGG network, which share exactly the same layer structure and parameters. Then, multi-scale features are extracted at each layer (represented in colors and they are stacked according to the scales). They are fused both within each image (in grey) into intra-image features and across images (in colors) into inter-image features, and then the intra-image features and the inter-image features are further fused at each scale. The features at different scales are represented in the corresponding colors. At last, the fused multi-scale features are summed up and upsampled to obtain the coarse co-segmentation results and final co-segmentation results.

Some methods utilize not only the low-level features but also the high-level semantic features from some pre-trained convolutional neural networks (CNNs). In the recent decade, CNNs have been successfully applied in various research areas and applications, such as image classification, image recognition [15], [16] and image semantic segmentation [17], [18]. In [4], [13], the similarities between regions are computed both by traditional low-level features and high-level CNN features. In [6], a deep descriptor transformation (DDT) method is proposed to evaluate the correlations of deep descriptors generated from pre-trained models and then locate the common category-consistent objects in a set of unlabeled images. In [14], the common objects are optimized to be separated from the backgrounds, to achieve the highest similarities among the objects and the lowest similarities between the objects and backgrounds, based on the high-level pre-trained CNNs

features.

There are also some methods that are fully supervised and trained based on deep CNNs. In [5], the networks are trained with the shared branches to extract intra-image features for each input image as well as a combined branch to extract inter-image features (co-features), and the two features are combined to final segmentation predictions.

In image co-segmentation area, deep CNNs have not been fully explored. Thus, in this paper, a novel neural network with multi-scale feature fusion is proposed. In this network, first, multi-scale convolutional features at multiple layers are extracted through the modified VGG network. Then, they are fused across images as the common inter-image features of the objects as well as within each image as intra-image features, and they are further fused to predict the coarse co-segmentation results. At last, the coarse co-segmentation results are refined via grab-cut based method to obtain the final segmentation results.

When training a neural network with multiple input images simultaneously, the cost of resources will increase rapidly with the inputs. To reduce the resource cost, the reduced input size, less downsampling and dilation convolution are jointly adopted in the proposed model.

Overall, the main contributions of the proposed model are twofold:

- A novel deep neural network with multi-scale feature fusion is proposed for image co-segmentation, and the experimental results demonstrate its comparable performance with the state-of-the-art methods.
- The proposed network has successfully reduced the cost of resources while achieving comparable results.

The rest of the paper is organized as follows. In section II, the details of the proposed model are explained. In section III, the experimental results are shown and analyzed, and the conclusions are drawn in section IV.

II. The proposed model

The proposed model overall contains three parts, as shown in Fig. 1, the first part is *a modified VGG network* (the left to the braces), the second part is *feature extraction and fusion* (in the braces) in which multi-scale features at multiple layers are extracted and fused, and the third part is *upsampling* (the right to the braces), in which losses are computed and the coarse co-segmentation results and final co-segmentation results are obtained.

For a group of input images $\{I_1, I_2, ...I_k, ..., I_K\}$, they are first passed through the modified VGG network, where K denotes the number of input images. Then, multi-scale features $F_k^{m,n}$, $m = \{1, 2, ..., 4\}, n = \{1, 2, ..., 5\}$ are extracted at multiple layers, where m denotes the scale and n denotes the layer. Then, $F_k^{m,n}$ are fused both within each image into intra-image features, denoted as ${F_{intra}}_k^m$, and across images into inter-image features, denoted as ${F_{inter}}^m$. And then ${F_{intra}}_k^m$ and ${F_{inter}}^m$ are further fused at each scale into co-features ${F_{co}}_k^m$. At last, ${F_{co}}_k^m$ are summed up into F_{cok} and upsampled to obtain the coarse co-segmentation results S_{cok} and final co-segmentation results R_{cok}.

A. The modified VGG network

The modified VGG network is modified from the standard VGG-16 network. We modify the parameters in some layers, and the modification details are shown in Table I, the parameters that are not mentioned remain the same as the original network.

The main purpose to modify these layers is that when input images are far small (such as 128×128), the feature blobs in the network should maintain their functions, thus, the pooling layers and the last two convolutional layers $Conv5\_2$ and $Conv5\_3$ are modified as Table I shows.

TABLE I
THE IMPLEMENTATION DETAILS OF THE LAYERS IN II-A.

layers	params	layers	params
Max pooling 3 Max pooling 4	Kernel: 3×3 Pad: 1 Stride:1	Conv5_2	Dilation: 2 Pad: 2 Output: 256
Conv4_1 Conv4_2 Conv4_3 Conv5_1	Output: 256	Conv5_3	Dilation: 5 Pad: 5 Output: 256

B. Feature extraction and fusion

As shown in Fig. 1, after images are passed through the modified VGG network, the multi-scale features ($scale1$, $scale2$, $scale3$, $scale4$) at multiple layers $F_k^{m,n}$ are obtained (features at $conv1\_2$, $conv2\_2$, $conv3\_3$, $conv4\_3$ and $conv5\_3$ are in cyan, pink, powder blue, green and blue, respectively) via dilated convolution. Then, they are fused at each scale within each image into intra-image features ${F_{intra}}_k^m$ and across the images into inter-image features ${F_{inter}}^m$. To make it clear in Fig. 1, the fusions in different scales are represented in different colors (arrows in purple, red, dark green, and cyan, respectively). And then, ${F_{intra}}_k^m$ and ${F_{inter}}^m$ are further fused and upsampled at each scale to obtain ${F_{co}}_k^m$. In this part, the structure of each image branch is the same and the parameters are shared. The details are shown in Table II.

C. Upsampling

In this part, the multi-scale co-features ${F_{co}}_k^m$ are summed up and upsampled to predict the coarse co-segmentation results S_{cok} and then to obtain the refined final co-segmentation results R_{cok}. For the loss function, we use the cross-entropy as the loss in each branch to train our network, as same as in [18]. The difference is that our model simultaneously computes K losses for K inputs correspondingly. In this part, the structure of each image branch is the same and the parameters are shared. The details are shown in Table III.

The coarse co-segmentation results S_{cok} may fail to obtain accurate object boundaries, due to the small input sizes of the training images. To refine the blurry S_{cok}, we use grab-cut method [19] to segment the coarse co-segmentation results.

TABLE II
THE IMPLEMENTATION DETAILS OF THE THE LAYERS IN II-B.

layers	params	layers	params
Scale1	Kernel: 3×3 Dilation: 2 Pad: 2 Output: 64	Fusion-intra Fusion-inter	Kernel: 1×1 Output: 256
Scale2	Kernel: 3×3 Dilation: 4 Pad: 4 Output: 64	Deconv1_1	Kernel: 1×1 Output: 256
Scale3	Kernel: 3×3 Dilation: 8 Pad: 8 Output: 64	Deconv1_2	Kernel: 3×3 Pad: 1 Output: 256
Scale4	Kernel: 3×3 Dilation: 16 Pad: 16 Output: 64	Upsampling1	Kernel: 2×2 Output: 256

Here we use four input images together in grab-cut for each time, and this helps to suppress some common background noise.

TABLE III
THE IMPLEMENTATION DETAILS OF THE LAYERS IN II-C

layers	params	layers	params
Deconv2_1	Kernel: 3×3 Pad: 1 Output: 128	Deconv3_2	Kernel: 3×3 Pad: 1 Output: 32
Upsampling2	Kernel: 2×2 Output: 128	Prediction	Kernel: 3×3 Pad: 1 Output: 2
Deconv3_1	Kernel: 3×3 Pad: 1 Output: 64		

From the implementation details we listed above in Tables I, II, and III, we can see that the cost of the proposed network has been effectively reduced and its volume (about $224.6M$ in total) is obviously much smaller than the standard-setting network. Nonetheless, the proposed network still achieves comparable performance.

III. EXPERIMENTS

A. Experimental settings

1) Datasets: To train our network, we use the Cosal2015 dataset [4] (50 image classes and 2015 images in total), the PASCAL-VOC dataset [20] (20 image classes and 1037 images in total) and the Coseg-Rep dataset [21] (23 image classes and 572 images in total) as the training dataset. To obtain the training image samples that each contains K images, we randomly rank the images belonging to a class, then select every K images in order as one training sample. In this paper, we set $K = 5$ and finally have 14496 training samples.

In order to verify the effectiveness of the proposed model, we choose the public image co-segmentation dataset [22] (38 image classes and 643 images in total, each class has 4 to 41 images) as the test dataset.

2) Implementation details: We implement our model based on the Caffe toolbox [23]. The proposed network is initialized randomly. In the training phase, we use the standard stochastic gradient descent (SGD) method with batch size 8, momentum 0.9 and weight decay 0.005. The learning rate is set to 10^{-6}. The input size is set to 128×128. The proposed model needs about 80k training iterations for convergence.

3) Evaluation metrics: To have an overall performance measurement, Intersection-over-Union (IoU) is used in our experiments.

B. Experimental results

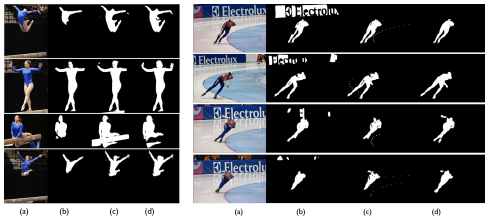

Fig. 3. Subjective evaluation results: (a) the original images, from (b) to (d), the results of [12], the results of the proposed model and ground truths, respectively.

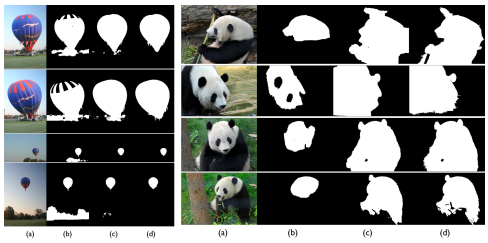

Fig. 4. Subjective evaluation results: (a) the original images, from (b) to (d), the results of [12], the results of the proposed model and ground truths, respectively.

We present the experimental results in both objective and subjective ways and compare the proposed model with a recent state-of-the-art model [12]. The objective evaluation results in terms of IoU are shown in Fig. 2. The proposed model obtains an average score of 0.711 on IoU, better than 0.702 on IoU by [12].

Fig. 5. Subjective evaluation results: (a) the original images, (b) the results of the proposed model, (c) ground truths.

Some subjective evaluation results are shown in Fig. 3 and Fig. 4. In Fig. 3, in the group of "skaters", the results of our model contain less background regions compared with

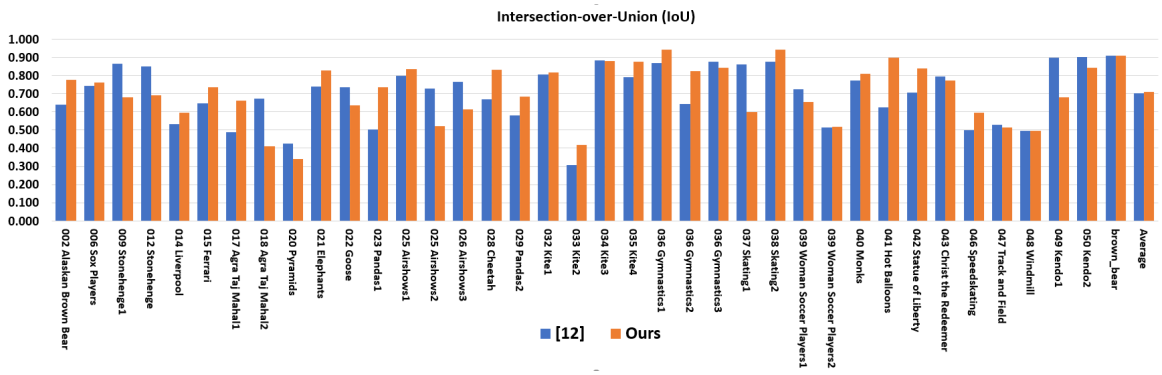

Fig. 2. Objective evaluation results.

the results of [12], while in "gymnastics" of Fig. 3, and in "balloons" and "panda" as shown in Fig. 4, our model segments the common objects more completely. We attribute these improvements to the successful utilization of deep network in which deep features are extracted.

There are also some limitations of the proposed model. In some cases, the common objects and noisy uncommon objects share similar high-level semantic features and differ in some low-level features such as color, and therefore it is difficult for the proposed model to distinguish them clearly, e.g. the group "woman soccer player 1", as shown in Fig. 5, the noisy objects are segmented as common objects. In the future work, the scheme to define co-features should be rethought to address this problem.

IV. Conclusion

In this paper, a novel deep neural network with multi-scale feature fusion is proposed for image co-segmentation, and it has successfully reduced the cost of the resources. The experimental results demonstrate its effectiveness for image object co-segmentation.

References

[1] Z. Liu, W. Zou, L. Li, L. Shen, and O. Le Meur, "Co-saliency detection based on hierarchical segmentation," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 88–92, 2014.

[2] L. Li, Z. Liu, W. Zou, X. Zhang, and O. Le Meur, "Co-saliency detection based on region-level fusion and pixel-level refinement," in *2014 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2014, pp. 1–6.

[3] L. Ye, Z. Liu, J. Li, W.-L. Zhao, and L. Shen, "Co-saliency detection via co-salient object discovery and recovery," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 2073–2077, 2015.

[4] D. Zhang, J. Han, C. Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2994–3002.

[5] L. Wei, S. Zhao, O. El Farouk Bourahla, X. Li, and F. Wu, "Group-wise deep co-saliency detection," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI Press, 2017, pp. 3041–3047.

[6] X.-S. Wei, C.-L. Zhang, J. Wu, C. Shen, and Z.-H. Zhou, "Unsupervised object discovery and co-localization by deep descriptor transformation," *Pattern Recognition*, vol. 88, pp. 113–126, 2019.

[7] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1943–1950.

[8] 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.

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

[10] 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.

[11] K. R. Jerripothula, J. Cai, and J. Yuan, "Quality-guided fusion-based co-saliency estimation for image co-segmentation and colocalization," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2466–2477, 2018.

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

[13] C.-C. Tsai, W. Li, K.-J. Hsu, X. Qian, and Y.-Y. Lin, "Image co-saliency detection and co-segmentation via progressive joint optimization," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 56–71, 2019.

[14] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Co-attention cnns for unsupervised object co-segmentation." in *International Joint Conference on Artificial Intelligence*, 2018, pp. 748–756.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[19] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3. ACM, 2004, pp. 309–314.

[20] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[21] J. Dai, Y. Nian Wu, J. Zhou, and S.-C. Zhu, "Cosegmentation and cosketch by unsupervised learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1305–1312.

[22] 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*. IEEE, 2010, pp. 3169–3176.

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.