SCIENCE CHINA
Information Sciences

• LETTER •

January 2023, Vol. 66 119101:1–119101:2
https://doi.org/10.1007/s11432-021-3515-6

Image co-segmentation based on pyramid features cross-correlation network

Jia CHEN[1], Yasong CHEN[1], Weihao LI[2], Zhi LIU[1,3], Sannyuya LIU[1,3] & Zongkai YANG[1,3*]

[1]*Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan* 430079*, China;*
[2]*Data61, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Canberra* ACT2601*, Australia;*
[3]*National Engineering Laboratory for Educational Big Data, Central China Normal University, Wuhan* 430079*, China*

Received 13 July 2021/Revised 26 October 2021/Accepted 17 March 2022/Published online 18 October 2022

Citation Chen J, Chen Y S, Li W H, et al. Image co-segmentation based on pyramid features cross-correlation network. Sci China Inf Sci, 2023, 66(1): 119101, https://doi.org/10.1007/s11432-021-3515-6

Dear editor,

Co-segmentation aims to segment objects with the same semantic information that simultaneously appears in two or multiple images. Vicente et al. [1] proposed the definition of object co-segmentation, i.e., "The task of jointly segmenting 'something similar' in a group or a pair of images is commonly referred to as co-segmentation". Co-segmentation pays attention to something that should be "objects" with the same semantics rather than other backgrounds in the images. For example, the common objects should be elements such as people, birds, cars, rather than background elements such as grass, curtain, sky. In contrast to semantic segmentation, co-segmentation employs common object information among images to produce segmentation. In co-segmentation, common information generally refers to the same semantic category, while the appearance and context of the common item in the image collection may be different. The co-segmentation task is challenging because it simply considers whether a given collection of images contains common objects.

Image co-segmentation methods can be divided into two categories [2]: image co-segmentation without deep learning and image co-segmentation based on deep learning. The deep learning method can extract high-level semantic features from images, considerably compensating for the flaws of hand-crafted features and completing the task of image co-segmentation more effectively. However, most deep learning-based co-segmentation algorithms to date have lacked depth mining of co-occurring information.

Unlike learning local features only by the convolutions, the attention mechanism [3, 4] uses a nonlocal mean filter to learn the relationship between long-distance pixels. The attention mechanism obtains the weight of all pixels by computing the similarity between each pixel and all global pixels, which has some noise reduction effect on irrelevant higher-level semantic information. From a channel standpoint, this attention method computes all channels at each pixel location as feature vectors. As a result, we developed a cross-correlation module that reduces the complexity while highlighting the location of common objects through the operation between channels.

Model framework. Our approach deeply explores the common semantic information between a pair of input images, thereby improving the segmentation effect of common objects. Figure 1 shows an overview of our framework, where C represents the cross-correlation operation based on pyramid features. The input is a pair of images I_A and I_B. It should be noted that during the training process, I_A and I_B are preprocessed to a uniform size of 513×513, and during the testing process, I_A and I_B can be in any size. I_A and I_B are processed by the Siamese encoder to obtain the corresponding high-level semantic feature maps f_A and f_B. Then, f_A and f_B are used as the inputs of C. C is the cross-correlation module. The output of light red C is F_A, and the output of light blue C is F_B. Then, we concatenate F_A and f_A, F_B and f_B, respectively, to get the high-level semantic feature maps as the input of the Siamese decoder, and finally the masks M_A and M_B are obtained by the Siamese decoder.

Cross-correlation module. Different channels represent different semantics in high-level semantic feature maps, which are normally independent, and the same class of objects will show higher activation values on the same channel [5]. Therefore, for common objects in the input image pair, they will show high activation values on the same channel. According to the receptive field concept, each position's activation value in a high-level feature map is associated with the original image's corresponding area. Therefore, the common objects in the image pair have three characteristics on the same channel. (1) They have the same semantic. (2) The positions of the active regions are different due to the varied placement of the common objects. (3) The sizes of the activated regions are different due to the arbitrary sizes of the image pair. To take advantage of characteristic (1), the

* Corresponding author (email: 13659885363@163.com)

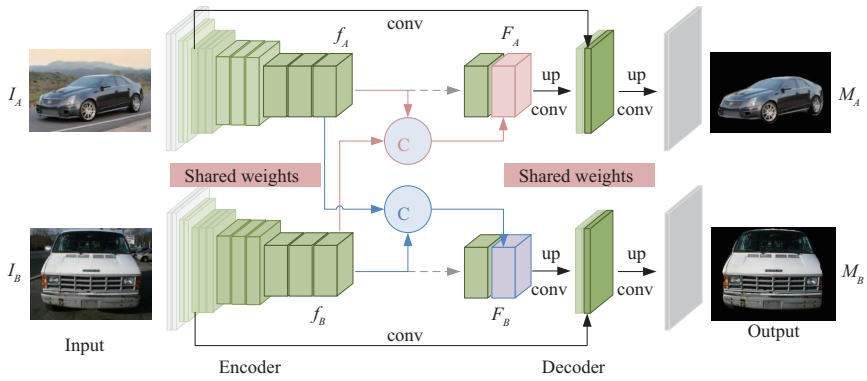

Figure 1 (Color online) Overview of our co-segmentation method based on pyramid features cross-correlation.

cross-correlation operation uses the depth-wise convolution for the high-level feature maps and performs the convolution operation on each channel. Channel fusion operation of common convolution does not exist in the depth-wise convolution. For characteristic (2), since the common objects may exist in any position of the image pair, the activated regions corresponding to the channel may appear in any position. In the cross-correlation operation, we regard the high-level feature maps of the first image as the bases and down-sample the second image to $k \times k$ pixels as the convolution kernel. Then, we use this $k \times k$ convolution kernel to convolve with the bases to compute the global correlation between the first image and the high-level features of the second image. For characteristic (3), the size inconsistency of the common objects is taken into consideration, and the feature pyramid is built by setting different subsampling sizes for the high-level semantic features of the convolution kernel to make the size of the activated regions on each channel of the two high-level semantic feature maps closer. When the centers of the two regions overlap, the correlation of the two regions may be higher, which helps to highlight the center position of the activated regions.

The cross-correlation operation based on the feature map f_A has the following mathematical description:

$$f'_A = \varphi_1(f_A), \tag{1}$$

$$S^i_B = \phi(\varphi_1(f_B); k_i), \tag{2}$$

$$\bar{f_A} = \frac{1}{m}\sum_{i=1}^{m} L_2(S^i_B) \otimes f'_A, \tag{3}$$

$$F_A = \varphi_2(\bar{f_A}), \tag{4}$$

where φ_1 is a 1×1 convolution layer without activation function; k_i is the size of the i-th pyramid feature map and $k_i \in \{11, 15, 19\}$; $L_2(\cdot)$ denotes L2 normalization; $\phi(\cdot; k_i)$ means to use bilinear interpolation to the feature maps and subsample them to the size $k_i \times k_i$; m is the number of pyramid layers; φ_2 is a 1×1 convolution layer with the activation function ReLU.

Additionally, we can use the above operational procedure to conduct the cross-correlation operation based on the feature map f_B by exchanging the two inputs of the cross-correlation module. The output of the cross-correlation operation is recorded as $F_B \in \mathbb{R}^{h_2 \times w_2 \times c}$. Then, we put F_A and F_B into their respective decoding branches to obtain the final segmentation results.

Conclusion. In this study, we propose an end-to-end deep learning method to accomplish image co-segmentation pair-wise. The Siamese encoder network is used to extract the high-level features. The core cross-correlation module is based on depth-wise convolution, which models the common semantic information between images from the perspective of feature similarity matching on each channel. And this module can highlight the center position of the high-level features of common objects. A multi-scale feature pyramid is constructed to improve the model's adaptability for objects of different sizes. We conducted the experiments on several public datasets. The experimental results show that our approach achieves state-of-the-art performance and can well accomplish the image co-segmentation task. Additionally, several groups of ablation experiments are designed to show the segmentation effect under different hyperparameters. The results show a good effect based on the cross-correlation operation of the pyramid features. Please see Appendixes A–C for details.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61605054, 62077017), Hubei Provincial Natural Science Foundation (Grant No. 2021CFB659), Fundamental Research Funds for the Central Universities (Grant Nos. CCNU22QN011, CCNU20TS032), and Science and Technology Innovation 2030 "New Generation Artificial Intelligence" Major Program (Grant No. 2020AAA0108804).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Vicente S, Rother C, Kolmogorov V. Object cosegmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011. 2217–2224
2 Xu H, Lin G, Wang M. A review of recent advances in image co-segmentation techniques. IEEE Access, 2019, 7: 182089–182112
3 Chen J, Chen Y, Li W, et al. Channel and spatial attention based deep object co-segmentation. Knowledge-Based Syst, 2021, 211: 106550
4 Chen H, Huang Y, Nakayama H. Semantic aware attention based deep object co-segmentation. In: Proceedings of Asian Conference on Computer Vision, 2018. 435–450
5 Li B, Wu W, Wang Q, et al. SiamRPN++: evolution of Siamese visual tracking with very deep networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 4282–4291