Deep Learning intra-image and inter-images features for Co-saliency detection

Min Li[1,2]
min.li8086@gmail.com

Shizhou Dong[3,4]
sz.dong@siat.ac.cn

Kun Zhang[5]
zhangkkk@whu.edu.cn

Zhifan Gao[6]
gaozhifan@gmail.com

Xi Wu[2]
wuxi@cuit.edu.cn

Heye Zhang[2]
heye.zhang@gmail.com

Guang Yang[7]
g.yang@imperial.ac.uk

Shuo Li[6]
slishuo@gmail.com

[1] Shaanxi Normal University
Xi'an, China

[2] Chengdu University of Information Technology
Chengdu, China

[3] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
Shenzhen, China

[4] Shenzhen College of Advanced Technology, University of Chinese Academy of Science
Shenzhen, China

[5] Wuhan University
Wuhan, China

[6] Western University
London, Canada

[7] Imperial College London
London, UK

Abstract

In this paper, we propose a novel deep end-to-end co-saliency detection approach to extract common salient objects from images group. The existing approaches rely heavily on manually designed metrics to characterize co-saliency. However, these methods are so subjective and not flexible enough that leads to poor generalization ability. Furthermore, most approaches separate the process of single image features and group images features extraction, which ignore the correlation between these two features that can promote the model performance. The proposed approach solves these two problems by multistage representation to extract features based on high-spatial resolution CNN. In addition, we utilize the modified CAE to explore the learnable consistency. Finally, the intra-image contrast and the inter-images consistency are fused to generate the final co-saliency maps automatically among group images by multistage learning. Experiment results demonstrate the effectiveness and superiority of our approach beyond the state-of-the-art methods.

Corresponding author: Heye Zhang (heye.zhang@gmail.com) and Xi Wu (wuxi@cuit.edu.cn)

1 Introduction

Unlike traditional saliency detection approaches [15, 22, 29, 43], co-saliency can explore the synergetic relationship from the given images group containing common attractive events to ameliorate characterizing co-saliency regions. Human visual system can locate common targets automatically among group images with same object and similar background [42]. Thus, co-saliency detection is an interesting research topic in computer vision community, which mimics human visual characteristics to detect the common foreground regions. Co-saliency detection has be applied in these fields including co-segmentation [7, 14, 33], weakly supervised learning[31, 38], video saliency detection [12, 23, 25], common pattern discovery [32, 37], etc. Besides, combining the depth information from RGBD images is a new co-salient study [10, 11].

Co-saliency detection is still a challenging research topic in computer vision because complicated background, object shelter, angle of observation and light condition changes could increase the difficulty of object detection. In order to detect co-salient events accurately, two issues should be concerned: 1) how to exact effective intra-image features and global consistency in a given images group; 2) how to apply the intra-image salient information and the global consistency information at group level to generate the final co-saliency maps. For the first issue, the intra image feature should reflect the unique characteristics of image events, and the consistency should reflect the correspondence relationship among images in a same group. For the second issue, a certain relationship between images with same foreground and common background should be built. This relationship can utilize the complementary information to remove the ambiguity of objects from the saliency regions in multiple images. Therefore, we should construct the global association between multiple images to extract consistency information for augmenting salient regions.

For accomplishing co-saliency detection among multiple images, some approaches have been proposed with different viewpoints in recent years [3, 27]. Among them, some of these methods rely on manual design metrics to extrct feature [13, 21]. However, these methods are subjective and dependent on researcher's prior knowledge, and thus they would loss some important information inevitably that can summarize the co-salient features better. Apart from that, they have focused on small and simple detection datasets [2, 20] with using low-level feature and fixed hand-disigned consistency computing mode. Therefore, these generalization ability needs to be enhanced. Some approaches [34, 41] focused on learning the co-saliency patterns. These methods extract high-level semantic features without detect the synergism between semantic features and global consistency [39, 41]. The inter images features of this method [34] is represented the integration of semantic features. However, the association between objects in a group is complex and abstract which can improve the results. So we proposed a new representation about the interaction relation among images in current group.

In this paper, we proposed a deep framework to discover the salient and interaction information at group level for co-saliency detection. Our framework focuses on more effective and robust feature representation to extract the intra-image contrast between salient objects and background and discover the relationship between the images in the same group simultaneously. Firstly, in the wake of increasing depth of the neural network layer, the intra-image contrast would be over dependent on the last layer, which would cause overfitting and poor transfer ability. To solve this problem, our work utilize the multistage features and pixel level detail information that be preserved at the same time. Thus, the feature representation consists of the image global information and the local detail information, which can make

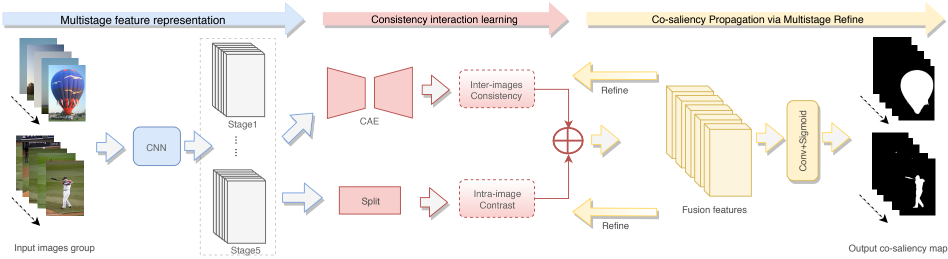


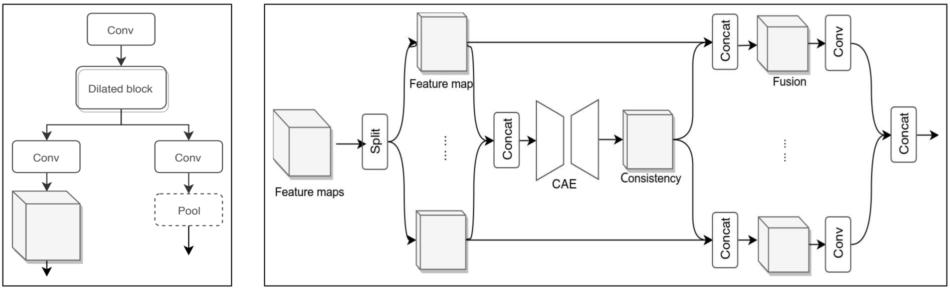
Figure 1: The framework of the proposed deep multistage learning method for co-saliency detection. Our method has three parts: 1) group inputs feature representation by multistage CNN, which can keep high-spatial resolution in deeper layer; 2) discover consistency and intra-image contrast respectively; and 3) co-saliency propagation via multistage learning to generate salieny maps.

the network more comprehensive and robust to improve the results in the end. To explore the inter images consistency, our work based on convolutional auto encoder (CAE) for group-wise interaction exaction can filter redundant and irrelevant information, and concurrently reserve the most useful features.

The main contributions of this work are summarized as follows: (1) We developed a deep co-saliency detection approach based on multi-scale semantics of single image and consistency of group images, which refines the process for feature endoding and co-saliency decoding. The approach incorporates the pyramidal hierarchical dialited convolution component to retain both the global features and detail information of group input images. Moreover, our method increases the receptive field in deeper layer and reduces the noise error during deconvolution. (2) For optimizing the learning process of intra-image contrast and inter-images consistency in conjunction, we set up a novel and expanded CAE formulation for co-saliency detecion, which takes advantage of the denosing capability from CAE to learn more robust expressions of group input and interaction relationships between group images. (3) We proposed a general and novel scheme via integrating the group images feature representation fine-grained refined stage-by-stage and the intrinsic relationship of intra-group images, which can achieve fantastical co-saliency maps. (4) We validated our method on three popular benchmarks. The experimental results show the superiority of our method to the state-of-the-art methods.

2 Related Work

The early existing methods were developed from image pairs. Jacobs *et al*. firstly defined co-saliency as the image saliency in the context of other images [18]. However, this definition has limitations that must use the same lens to get the similar image pairs for detecting. This limitation restricts the application of co-saliency without universality. Afterwords, Li *et al*. proposed a linear combination of the single-image saliency map and the multi-image saliency map to generate co-saliency map [20]. Furthermore, they established the first public dataset for co-saliency. Then, Chang *et al*. established a unsupervised model via MRF optimization algorithm that has an energy function [5]. To solve the problem of co-saliency detection, Chen *et al*. presented sparse distribution representation without correspondence matching to enhance the proformance [6]. To expand the co-saliency detection problem from two



(a) (b)

Figure 2: (a) The illustration of the multistage CNN backbone (dotted line represents only using pool layer in the first two phases); (b) The paradigm of the conversion of group features shape for self-adapive extract the fusion of consistency and intra-image contrast.

images to multiple images, Li *et al*. designed three types of visual descriptors, and used minimum spanning tree to determine the image matching order [21]. Fu *et al*. proposed an approach based cluster for generate the final co-saliency maps by three visual attention cues [13]. These approaches need a lot of prior knowledge and rely on low-level features, such as color, texture, edge. They were designed share a certain pixel-level consistency in bottom-up pathway and without object-level feature representation. Consequently, It is easily falling into local optimum rather than global optimum results.

Different from the bottom-up methods, some methods utilize multiple saliency submaps predicted by existing methods and fuse them to integrate the co-saliency maps. Huang *et al*. constructed a multiscale superpixel pyramid and utilized joint low-rank analysis to obtain an adaptively fused saliency map [17]. Cao *et al*. proposed to use rank-one constraint to combine the saliency maps [3]. They modified their model by utilizing low rank matrix approximation and low rank matrix recovery to formulate the general consistency criterion [4]. These methods weight these saliency submaps cues as mid-level features, which are based on existing algorithms and fusion to discover co-saliency cues. Recently, Zhang *et al*. applied CNN to extract the high-level semantic features and disigned a framework with self-paced multiple-instance learning [40]. In addition, Zhang *et al*. proposed another way which use domain adaptive CNN with transfer RBM layers [39]. Rather than estimating the intra image information and the consistency respectively, Wei *et al*. inferred the co-salient regions by a framework based FCN [34]. As can be seen from the above part, although the existing works completed the co-saliency detection task, there still existed some debatable parts need discussed and improved.

Furthermore, this paper proposes the approach to detect co-salient objects based on CNN that can balance the local and global features and has well robustness via the modified CAE and multistage learning. This approach aims at expanding spacings among classes and reducing intra-class spacings. Thus, we combine different stages parameters to enhance performance by the relevance among stages.

3 Methodology

Our co-saliency detection method can be summarized as a high-level overview by Figure 1. This scheme is an end to end deep neural network based on saliency block and CAE. As part in saliency block, the dilated convolutional layers are skip connected with each one to observe features. These features can describe the images with various-level representation to be reused in the next process. The modified CAE part is utilized to discover the inter-images features or consistency. Modeling these features, the final classifier predicts the more precise co-saliency maps group.

Specially, we develop the following modifications to the original structure. On the one side, instead of the traditional techniques like Fully Convolutional Networks (FCN), our approach has five stages that observe multiple semantic feature around dilated block in every saliency block. We combine not only different stage features, but also loss function from each stage. On the other side, we are the first to discover the coordination relation between images by designing this special modified CAE. We describe the detailed improvements of our network.

3.1 Features Representation via High-spatial Resolution CNN

We propose a deep multistage CNN that can retain the high spatial resolution in deeper layers to extract salient object feature for each image. We were inspired by DenseNet proposed by Huang *et al*. due to its high efficiency [16]. By highly reusing features, we can achieve better convergence while reducing the amount of parameters and calculations.

Intra-image Contrast

In co-saliency detection task, due to objects have variant scales, designing an effective and robust CNN that can adapt different objects scales is a great challenge. In order to solve this problem, we apply saliency blocks to extract single image features. Although recent feature pyramid networks [26] can work on multiple scales, due to it relys on downsampling layer or stride to expand the receptive field for detecting the large targets at the expense of spatial resolution. In particular, different from the traditional feature pyramid algorithm, we maintain not only the high resolution feature maps, but also large receptive field by employing a dilated encoder-decoder structure for images representation. As shown in Figure 2(a), a saliency block has two branches. We only use the pooling layer in the first two stages to change the feature map scale, which means the final feature map can better balance the image global information and local details. At each stage, the current stage concatenates the previous stages feature maps, which can maintain both semantics and details. Finally, the output of stage5 is regarded as the intra-image contrast.

3.2 Channel-wise Interaction Representation

We propose a channel-wise expanded CAE to accelerate co-saliency detection effect simultaneously with avoiding the error caused by small batch size. The channel-wise expanded CAE can use convolutional operation to calculate the sum of nonlinear superposition of signals. In the encoding and decoding processing, a group images feature can be remodeled and minimum the signal reconfiguration error. Unlike the previous CAE, we did some extensions and improvements for learning the consistency of group images. As shown in Figure 3, we proposed a new fusion strategy about group images consistency, which can convey well collaboration or synergistic feature and common attributes between images in a group input. By building this modified structure, the parameters of the method can be better optimized without being affected by limited data.

Inter-images Consistency

we first describe a general formulation of the inter-images feature representation, and then present our approach improvements in this formulation. Generally speaking, the con-

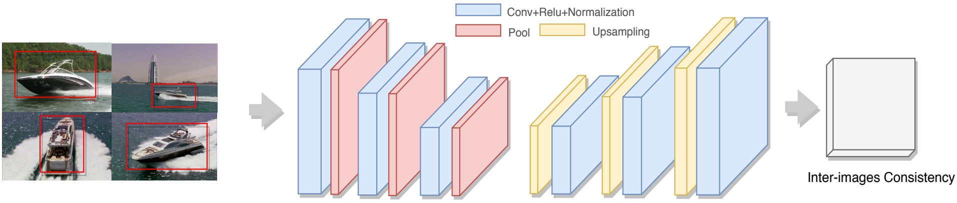


Figure 3: The modified CAE for exploring the joint aggregation.

sistency can be defined as in learning method:

$$x_{co} = f_{cosal}(X, \theta_{cae}) \tag{1}$$

where x_{co} is the consistency which represent the interaction relationships of inter-images in the current group and the X is the feature maps group represented by Sect. 3.1. f_{cosal} is the modified CAE process in Figure 3. This bottleneck structure is able to keep the most kernel and joint information from images group, which filter out the redundant and noise information.

Channel-wise Representation

A group image features are constructed to represent the consistency by our redisigned CAE structure. As show in Figure 2(b), we have some conversion on group features dimension to enable the intra-group consistency to be learned. Specifically, the group features input tensor is in $(G, H, W, 1)$ order, then we reshape it to $(1, H, W, G)$ order for constructing the intra-group consistency by CAE, where G is the input group images number, H and W are the images height and width, the feature channel is 1. More importantly, we present normalization in this tensor order. It can be preformed as the formula next:

$$\hat{x}_{co_i} = \frac{x_{co_i} - \mu_i}{\sqrt{\sigma_i^2 + \varepsilon}} \tag{2}$$

where x_{co_i} is the intra-group consistency at index i, μ and σ are the mean and variance along the $(1, H, W)$ axes within each image. This modified structure can increase the robustness and generalization ability and avoid overfitting.

3.3 Co-saliency Propagation via Multistage Learning

By sharing representation among group images, we can generalize a better solution upon co-saliency detection task. Typically, we pay attention to a certain benchmark for training model and fine tuning network. Multistage learning is viewed as a form of interactive transfer, which introduces an interactive relation to increase model. Furthermore, the single stage can't utilize correlation and attributes between different stages to improve precision. In the processing of training, the final results too rely on the last layer parameters, which could leads to overfitting. In contrast to single stage learning, we optimize more than one loss function. By sharing hidden layers parameters, our method increases performance effectively due to implicit data argumentation and concentrating on key features. In addition, although the batch data flow back-propagation incorporate the similarity of intra-group images and the discrepancy of inter-group images, the network also could gain by further enhance. In this section, we develop an end-to-end framework to detect co-salient objects within groups,

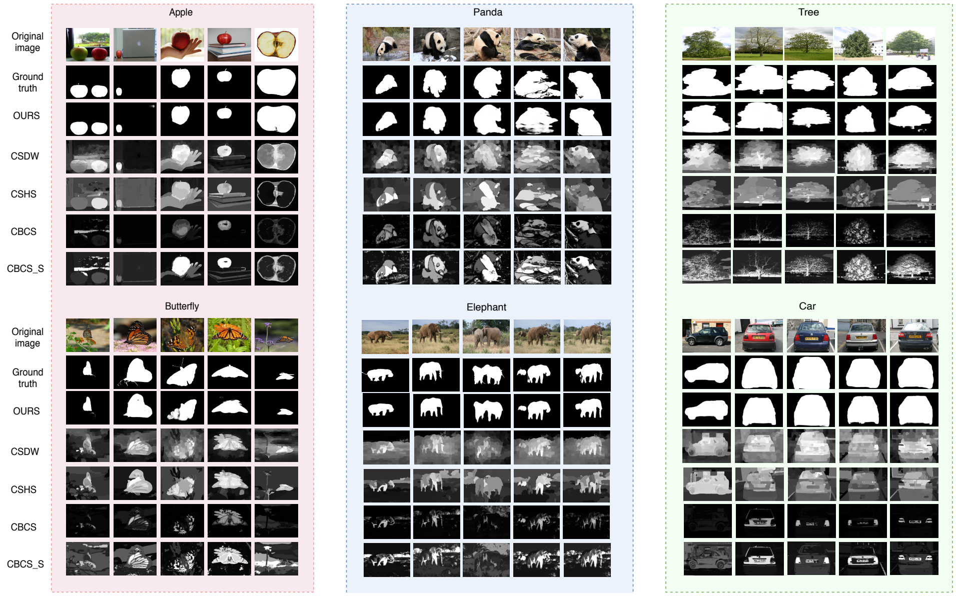


Figure 4: Visual comparsion of co-saliency detection from different groups on three benchmark. The pink block, blue block, and the green block are from the Cosal2015 dataset, the iCoseg dataset, and the MSRC dataset. Apparently, our approach has a well performance.

where the single features and the joint feature from the current group are refined with multistage learning. Specifically, we first minimize a fusion lost function to optimize the deep neural network. Let $X = x_m, m = 1, ..., M$ and $G = g_m, m = 1, ..., M$ denote the training images group and saliency maps group. The function is defined as:

$$L_{final}(\vartheta^{(n)}, \theta_{inter}, \theta_{cosal}) = \sum_{m=1}^{M} \sum_{n=1}^{N} l_{fuse}(g_m | x_m; \vartheta^{(n)}, \theta_{inter}, \theta_{cosal}) \tag{3}$$

where $\vartheta^{(n)}$ is the parameters of the nth intra-image feature extraction stage and l_{fuse} denotes the sum of the pixl-level mean absolute errors loss, the weighted cross-entropy loss [30] and the dice loss [27]. The θ_{inter} and θ_{cosal} are the parameters of the consistency structure and layers which flexibly merge the intra-image features and the inter-images consistency. By utilizing the relevance among each stage, our method can enlarge the inter-class distances and decrease intra-class distances.

4 Experiments and Results

4.1 Experiment Setup

Datasets

We evaluate the performance of our method on three popular benchmark datasets that are publicly avaiable: iCoseg [2], MSRC [35] and Cosal2015 [41]. The iCoseg dataset contains 38 image groups of totally 643 images and it is a challenging dataset to evaluate co-saliency

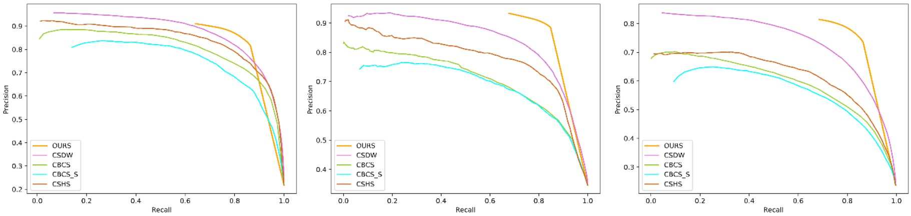


Figure 5: Comparsion results on the three benchmark datasets. From left to right, the PR curve of iCoseg, MSRC, and Cosal2015.

detection techniques or methods because the images collected by this dataset contain complex backgrounds and multiple co-salient objects. MSRC is another dataset which is widely used to evaluate the co-saliency detection approaches. The MSRC consists of 8 groups of totally 240 images and it contains bicycle and other objects which are difficult to detection and segmentation. We did't use the grass images because of these images have no co-salient objects. A newly published and highly cluttered dataset Cosal2015 has 50 groups and totally 2015 images which are collected from ILSVRC2014 detection benchmark and the YouTube video set. These datasets all have pixel-wise ground-truth labels and hand annotations. Compared to the previous datasets, the Cosal2015 is bigger and more challenging for evaluating co-saliency detection algorithm.

Evaluation Metrics

We evaluate the performance of existing methods and our method in the experiments via comparing ground-truth (G) with the co-saliency map (C). We utilize five criterias for qualitative and quantitative experiments, including the Precision score, Recall score, Precision-Recall (PR) curve, F-measure, and average precision (AP) score. In the field of co-saliency detection, researchers always first segment co-saliency maps via 256 thresholds from 0 to 255 [24, 34, 41]. The Precision and Recall (PR) curve is drawn by using the precision rate and the recall rate at 256 thresholds, the higher the better. Moreover, the average precision (AP) score is the area under the PR curve. Specifically, the precision and the recall are defined as:

$$Precision = \frac{sum(C,G)}{sum(C)}, \quad Recall = \frac{sum(C,G)}{sum(G)} \tag{4}$$

where $sum(C,G)$ is the sum of corresponding pixels multiply, $sum(C)$ is the sum of the co-saliency map all pixels values, $sum(G)$ is the sum of the ground-truth all pixels values.

In the experiments, the F-measure based on the precision and recall is defined as:

$$F\text{-}measure = \frac{(1+\beta^2)\cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \tag{5}$$

where β^2 is always set to 0.3 suggested by [34, 40, 41] and using an adaptive threshold was proposed in [19].

Implementation Details

In the experiment, our co-saliency approach is implemented by using TensorFlow toolbox. Our network is initialized by a pretrained version that use the single image to predict saliency map. In addition, we transfer the model for exploring inter-images interaction relation and enhancing the performance by fine-tuning. Moreover, our training dataset includes MSRA 10k [9], THUR-15K [8], and DUT-OMRON [36]. We resized all the images to

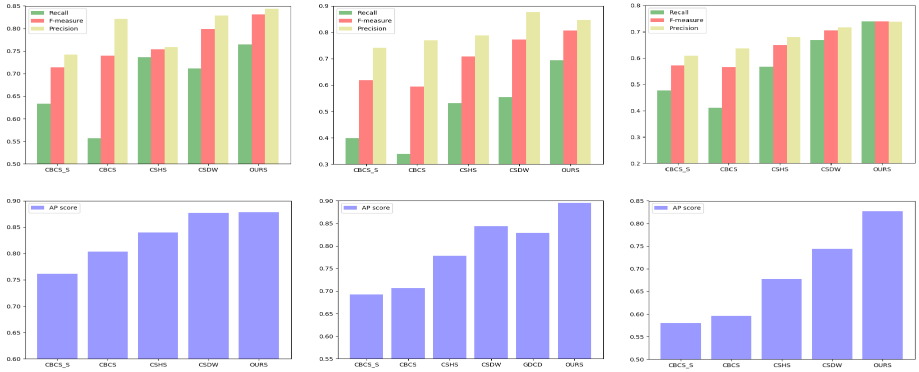


Figure 6: Quantitative comparisons (AP score, Precision, Recall and F-measure) between the proposed method and the state-of-the-art methods on three benchmark datasets. From left to right, the metric from iCoseg, MSRC, and Cosal2015. Notice that the method GDCD [34] has no open source codes or map results, we have to quote their AP score (only providing the MSRC dataset score). Obviously, our method achieves the best performence.

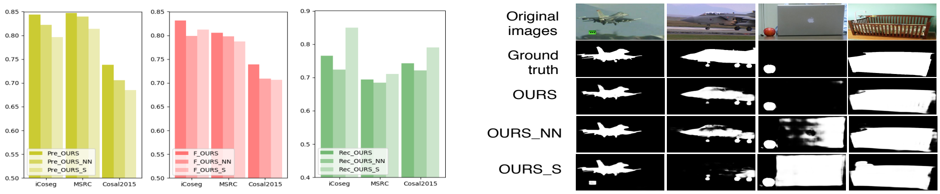
256x256 pixels and used stochastic gradient descent with momentum parameter 0.9, the learning rate is set to 0.000002, and the weight decay is 0.0001. We need about 80 epochs for training.

4.2 Comparsion with State-of-the-Art Methods

In this section, we compared the developed method with four state-of-art methods, i.e. CSDW [41], CSHS [28], CBCS [13], CBCS_S [13]. In order to evaluate each method subjectively, some examples in each dataset and method are shown in Figure 4. The co-saliency maps are genereted by the above methods and ours. As shown by the examples from iCoseg dataset, i.e., the images groups of panda and elephant. Some of these images have more than one co-saliency events or shelter, which raises issue or bar about co-salient detection. The images groups of apple and butterfly from Cosal2015 dataset have noise background or object to increase detection difficulty. These examples indicate the robustness of algorithms. It can be concluded from these maps that performance gain of our method, which is much better than the others. Besides, we show the different metrics results for quantitative comparison. The PR curves is shown in Figure 5. The AP scores and Precision, Recall F-measure using adaptive threshold are shown in Figure 6. Note that the AP score of our method (0.8781) is higher than CSDW (0.8766) with signficant difference ($p\text{-value}<0.05$ in F-test). As can be seen these figures, our method can reach or exceed the previous best method, i.e., CSDW. Consequently, the proposed method validity and accuracy are proved.

4.3 Model Analysis

In this section, we chose different experiments manners to analyse the structures influence in the developed framework. The results can be shown in Figure 7, which indicates that the single saliency network (OURS_S) is easy to interfere by all kinds of noise. Because precision is more important than recall in co-saliency task [1], the method merged different features but without special normalization (OURS_NN) is better than the OURS_S. How-



(a) (b)

Figure 7: (a) Evaluation of the Precision, Recall, and F-measure of three structures in three datasets. (b) Some examples to show that our approach efficiency and comprehensiveness.

ever, OURS can still enhance the preformance. The reason is that our method considers not only exploring the interaction by multistage learning, but also improving stability by special normalization. Compared with the OURS_S and OURS_NN, the proposed method can improve resistance to interaction and detection rate efficiently.

5 Conclusion

In this paper, we have proposed a novel end-to-end deep co-saliency detection method which self-adaptively learns the correspondence constraint based on the image group as supplementary information to solve the co-salient problem. In addition, by extracting feature from image, we proposed a multistage CNN with high-spatial resolution in deep layers to learn and transfer the feature representation. By using the modified CAE with special normalization to explore the interaction relation between images in the current group. Compared with the state-of-the-arts, comprehensive experments results indicate our method effectiveness.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1597–1604, 2009.

[2] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen. icoseg: interactive co-segmentation with intelligent scribble guidance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3169–3176, 2010.

[3] Xiaochun Cao, Zhiqiang Tao, Bao Zhang, Huazhu Fu, and Xuewei Li. Saliency map fusion based on rank-one constraint. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2013.

[4] Xiaochun Cao, Zhiqiang Tao, Bao Zhang, Huazhu Fu, and Wei Feng. Self-adaptively weighted co-saliency detection via rank constraint. *IEEE Transactions on Image Processing*, 23(9):4175–4186, 2014.

[5] Kai Yueh Chang, Tyng Luh Liu, and Shang Hong Lai. From co-saliency to co-segmentation: an efficient and fully unsupervised energy minimization model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2129–2136, 2011.

[6] Hwann Tzong Chen. Preattentive co-saliency detection. In *IEEE International Conference on Image Processing (ICIP)*, pages 1117–1120, 2010.

[7] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Enriching visual knowledge bases via object discovery and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2035–2042, 2014.

[8] Ming Ming Cheng, Niloy J. Mitra, Xiaolei Huang, and Shi Min Hu. Salientshape: group saliency in image collections. *Visual Computer International Journal of Computer Graphics*, 30(4):443–453, 2014.

[9] Ming Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip Torr, and Shi Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[10] Runmin Cong, Jianjun Lei, Huazhu Fu, Weisi Lin, Qingming Huang, Xiaochun Cao, and Chunping Hou. An iterative co-saliency framework for rgbd images. *IEEE Transactions on Cybernetics*, In Press, 2018. doi: 10.1109/TCYB.2017.2771488.

[11] Runming Cong, Jianjun Lei, Huazhu Fu, Qingming Huang, Xiaochun Cao, and Chunping Hou. Co-saliency detection for rgbd images based on multi-constraint feature matching and cross label propagation. *IEEE Transactions on Image Processing*, 27(2): 568–579, 2018.

[12] Yuming Fang, Zhou Wang, and Weisi Lin. Video saliency incorporating spatiotemporal cues and uncertainty weighting. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2013.

[13] Huazhu Fu, Xiaochun Cao, and Zhuowen Tu. Cluster-based co-saliency detection. *IEEE Transactions on Image Processing*, 22(10):3766–3778, 2013.

[14] Huazhu Fu, Dong Xu, Stephen Lin, and Jiang Liu. Object-based rgbd image co-segmentation with mutex constraint. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4428–4436, 2015.

[15] Kuang Jui Hsu, Yen Yu Lin, and Yung Yu Chuang. Weakly supervised saliency detection with a category-driven map generator. In *British Machine Vision Conference (BMVC)*, 2017.

[16] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Weinberger. Densely connected convolutional networks. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[17] Rui Huang, Wei Feng, and Jizhou Sun. Saliency and co-saliency detection by low-rank multiscale fusion. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2015.

[18] David E. Jacobs, Dan B. Goldman, and Eli Shechtman. Cosaliency: where people look when comparing images. In *ACM Symposium on User Interface Software and Technology (UIST)*, pages 219–228, 2010.

[19] Yangqing Jia and Mei Han. Category-independent object-level saliency detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1761–1768, 2013.

[20] Hongliang Li and King Ngi Ngan. A co-saliency model of image pairs. *IEEE Transactions on Image Processing*, 20(12):3365–3375, 2011.

[21] Hongliang Li, Fanman Meng, and King Ngi Ngan. Co-salient object detection from multiple images. *IEEE Transactions on Multimedia*, 15(8):1896–1909, 2013.

[22] Jian Li, Martin Levine, Xiangjing An, and Hangen He. Saliency detection based on frequency and spatial domain analyses. In *British Machine Vision Conference (BMVC)*, 2011.

[23] Wei-Te Li, Haw-Shiuan Chang, Kuo-Chin Lien, Hui-Tang Chang, and Yu-Chiang Frank Wang. Exploring visual and motion saliency for automatic video object extraction. *IEEE Transactions on Image Processing*, 22(7):2600–2610, 2013.

[24] Xi Li, Yao Li, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Contextual hypergraph modeling for salient object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3328–3335, 2014.

[25] Yong Li, Bin Sheng, Lizhuang Ma, Wen Wu, and Zhifeng Xie. Temporally coherent video saliency using regional dynamic contrast. *IEEE Transactions on Circuits Systems for Video Technology*, 23(12):2067–2076, 2013.

[26] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017.

[27] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):353–367, 2011.

[28] Zhi Liu, Wenbin Zou, Lina Li, Liquan Shen, and Olivier Le Meur. Co-saliency detection based on hierarchical segmentation. *IEEE Signal Processing Letters*, 21(1):88–92, 2014.

[29] Qinmu Peng, Yiu Ming Cheung, Xinge You, and Yuan Yan Tang. A hybrid of local and global saliencies for detecting image salient region and appearance. *IEEE Transactions on Systems Man Cybernetics : Systems*, 47(1):86–97, 2017.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[31] Parthipan Siva, Chris Russell, Tao Xiang, and Lourdes Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3238–3245, 2013.

[32] Hung-Khoon Tan and Chong-Wah Ngo. Common pattern discovery using earth mover's distance and local flow maximization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1222–1229, 2005.

[33] Le Wang, Gang Hua, Rahul Sukthankar, Jianru Xue, and Nanning Zheng. Video object discovery and co-segmentation with extremely weak supervision. In *European Conference on Computer Vision (ECCV)*, pages 640–655, 2014.

[34] Lina Wei, Shanshan Zhao, Omar El Farouk Bourahla, Xi Li, and Fei Wu. Group-wise deep co-saliency detection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3041–3047, 2017.

[35] Jonh M Winn, Antonio Criminisi, and Thomas P Minka. Object categorization by learned universal visual dictionary. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1800–1807, 2005.

[36] Chuan Yang, Lihe Zhang, Huchuan Lu, Ruan Xiang, and Ming Hsuan Yang. Saliency detection via graph-based manifold ranking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3166–3173, 2013.

[37] Junsong Yuan and Ying Wu. Spatial random partition for common visual pattern discovery. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

[38] Dingwen Zhang, Junwei Han, Gong Cheng, Zhenbao Liu, Shuhui Bu, and Lei Guo. Weakly supervised learning for target detection in remote sensing images. *IEEE Geoscience Remote Sensing Letters*, 12(4):701–705, 2014.

[39] Dingwen Zhang, Junwei Han, Chao Li, and Jingdong Wang. Co-saliency detection via looking deep and wide. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2994–3002, 2015.

[40] Dingwen Zhang, Deyu Meng, Chao Li, Lu Jiang, Qian Zhao, and Junwei Han. A self-paced multiple-instance learning framework for co-saliency detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 594–602, 2015.

[41] Dingwen Zhang, Junwei Han, Chao Li, Jingdong Wang, and Xuelong Li. Detection of co-salient objects by looking deep and wide. *International Journal of Computer Vision*, 120(2):215–232, 2016.

[42] Dingwen Zhang, Huazhu Fu, Junwei Han, Ali Borji, and Xuelong Li. A review of co-saliency detection algorithms: Fundamentals, applications, and challenges. *ACM Transactions on Intelligent Systems Technology*, 9(4):1–31, 2018.

[43] Wenbin Zou, Kidiyo Kpalma, Zhi Liu, and Joseph Ronsin. Segmentation driven low-rank matrix recovery for saliency detection. In *British Machine Vision Conference (BMVC)*, 2013.