Group-wise Deep Co-saliency Detection

Lina Wei[1], **Shanshan Zhao**[1], **Omar El Farouk Bourahla**[1], **Xi Li**[1,2,*], **Fei Wu**[1]
[1] Zhejiang University, Hangzhou, China
[2] Alibaba-Zhejiang University Joint Institute of Frontier Technologies, Hangzhou, China
{linawzju, zsszju, obourahla, xilizju, wufei}@zju.edu.cn

Abstract

In this paper, we propose an end-to-end group-wise deep co-saliency detection approach to address the co-salient object discovery problem based on the fully convolutional network (FCN) with group input and group output. The proposed approach captures the group-wise interaction information for group images by learning a semantics-aware image representation based on a convolutional neural network, which adaptively learns the group-wise features for co-saliency detection. Furthermore, the proposed approach discovers the collaborative and interactive relationships between group-wise feature representation and single-image individual feature representation, and model this in a collaborative learning framework. Finally, we set up a unified end-to-end deep learning scheme to jointly optimize the process of group-wise feature representation learning and the collaborative learning, leading to more reliable and robust co-saliency detection results. Experimental results demonstrate the effectiveness of our approach in comparison with the state-of-the-art approaches.

1 Introduction

In principle, co-salient object detection [Batra *et al.*, 2010; Chang *et al.*, 2011; Li and Ngan, 2011; Fu *et al.*, 2013; Li *et al.*, 2013] is defined as the problem of discovering the common and salient foregrounds from an image group containing multiple images at the same time. It has a wide range of applications on computer vision tasks, such as image or video co-segmentation [Fu *et al.*, 2015b; 2015a; Wang *et al.*, 2015], object localization [Tang *et al.*, 2014; Cho *et al.*, 2015], and weakly supervised learning [Siva *et al.*, 2013].

In order to detect co-salient regions precisely, we need to focus on two key points: 1) how to extract effective features to represent the co-salient regions; 2) how to model the interactive relationship between images in a group to obtain the final co-saliency maps. For 1), feature representation in the co-saliency detection task should not only reflect the individual properties of each image itself, but also express the relevance and interaction between group images. For 2), we know that images within a group are contextually associated with each other in different ways such as common objects, similar categories, and related scenes. The co-saliency detection job tries to use this information to find the target saliency maps, so we can utilize the consistency information within these image groups and capture an interaction between the images so that they mutually reinforce and enhance each other's saliency regions.

For tackling lots of challenges, we need to design a model that can extract robust features that reflect the individual properties of each image as well as features that represent the group-wise information such as group consistency, object interactions and, to a minor extent, the objects that are present in only single images but not the rest of the images. A series of approaches have been proposed from different points of view. Some methods [Chang *et al.*, 2011; Fu *et al.*, 2013; Li *et al.*, 2013; Cheng *et al.*, 2014] consider that the co-salient objects appearing in the group images should share a certain consistency in both low-level feature and high-level semantic feature [Zhang *et al.*, 2016b; 2015; 2016a], however, they do not model the interaction between the group-wise features and single image features, which can contain information that can improve the results. Some approaches detect the single-image individual saliency and the common salient regions of a group in a separate manner [Ge *et al.*, 2016; Li *et al.*, 2013] and, they also detect the intra-image and inter-image saliency separately from other information priors, such as the objectness [Li *et al.*, 2014; Liu *et al.*, 2014], the center priors [Chen and Hsu, 2014], and the border connectivity [Ye *et al.*, 2015]. Usually, calculating the intra-image and inter-image saliency separately is incapable of well capturing the intrinsic semantic interaction information among images within each group, which is important to the co-saliency detection quality.

Motivated by this observation, we propose a co-saliency deep model based on a fully convolutional network(FCN) with group input and group output. Our aim is to make use of all the information available and create a robust and effective network. Our model needs to take into account both the image properties and the intra group information while processing the co-saliency results. We design our network to be fully convolutional, this allows it to fully benefit from the

*corresponding author

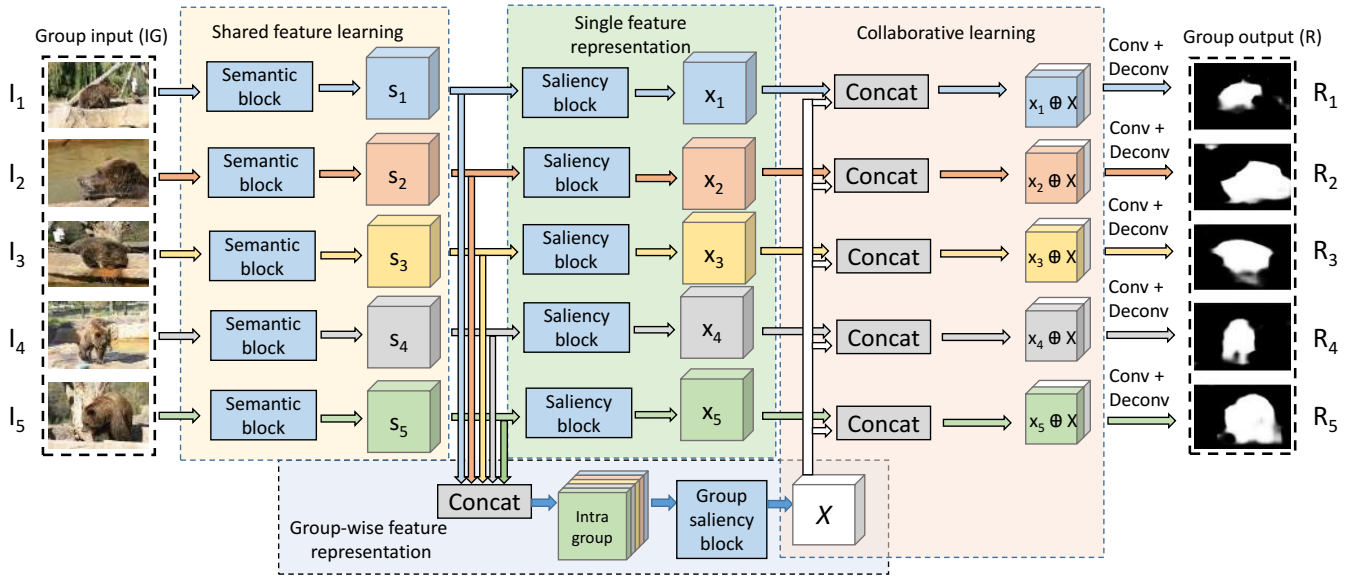

Figure 1: Illustration of the proposed network architecture for co-saliency detection. The group images $\{I_1, I_2, ..., I_5\}$ first go through the semantic feature extraction block, its results which are the image features $\{s_1, s_2, ..., s_5\}$, as well as a concatenated version of them are then passed through the (group) saliency feature extraction block, then, the result X coming from the concatenated version is further concatenated to the results coming from each of the other individual results$\{x_1, x_2, ..., x_5\}$, then a convolution + deconvolution is applied on each to recover the saliency. $\bigoplus$ in this figure means concatenate operation

local relationships between the pixels in an image, it is also designed deep enough to have a large receptive field. The network will extract the semantic features of the images, then will be divided into two branches. Namely, one processes each image individually and the other takes into account all the image group, the branches are later merged. This allows the network to learn features not only from the individual image properties, but also from the intra group properties, leveraging the shared and unique information between the images, resulting in accurate co-saliency maps. Our deep model takes a data-driven learning pipeline for capturing the collaboration and consistency intra image group, and is trained end-to-end.

The main contributions of this work are summarized as follows:

First, we propose a unified group-wise deep co-saliency detection approach with group input and group output, which takes advantage of the interaction relationships between group images. The proposed approach performs feature representation for both single image(e.g., individual objects and unique properties) and group consistency (e.g., common background and similar foreground), which generally leads to an improvement in the performance of the co-saliency detection.

Second, we set up an end-to-end deep learning scheme(FCN) to jointly optimize the process of group-wise feature representation learning and the collaborative learning, leading to more reliable and robust co-saliency detection results. The collaborative learning process combines the group-wise saliency and single image saliency in a unified framework that model a better interaction relationships between group images.

2 Proposed Approach

2.1 Problem Formulation

Given a group of images $IG = \{I_i\}_{i=1}^K$ where K is the number of images in group. Our goal is to discover the co-salient regions $R = \{R_i\}_{i=1}^K$ for this image group, where R_i is the saliency region for image I_i. In a simple saliency problem, each saliency region depends on its image and so we theoretically wish to find R_i such that :

$$R_i = f(I_i; \Theta) \tag{1}$$

where f is a regression function that takes the image I_i as input, and outputs the desired saliency map by learning a set of parameters Θ. However, in our case, I_i exists within a group of images that are contextually associated with each other, so that each saliency region has a certain interaction and depends on that of the other images. This changes the function we want to find to :

$$R = F(IG; \Theta) \tag{2}$$

In order to formulate the framework, we propose an end-to-end FCN with group input and group output which will process all the images at the same time and combine them at the feature level covering the needed theoretical necessity of taking into account all the images. The proposed group-wise co-saliency detection approach mainly consists of two components: 1) encoding the group images into co-feature representations by group-wise and single image feature learning to better obtain the comprehensive information, 2) collaborative learning by combing the group-wise feature with the single image feature through a unified joint learning structure which can comprehensively preserve common objects of the group

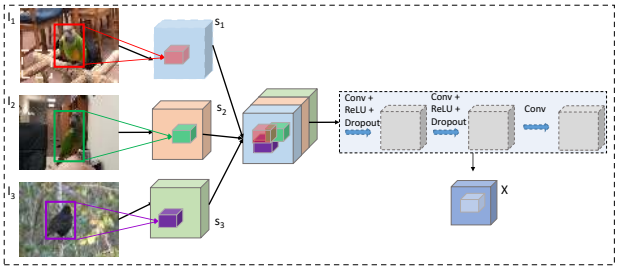

Figure 2: Illustration of the working mechanism of group-wise feature learning by fusion strategy. The bounding boxes with different colors indicate the birds appear in different images and they are correlated with each other. We solve the problem by three stages: 1) group wise consistency investigation by group feature interaction; 2) joint aggregation by feature concatenation; and 3) saliency computation by a convolutional process.

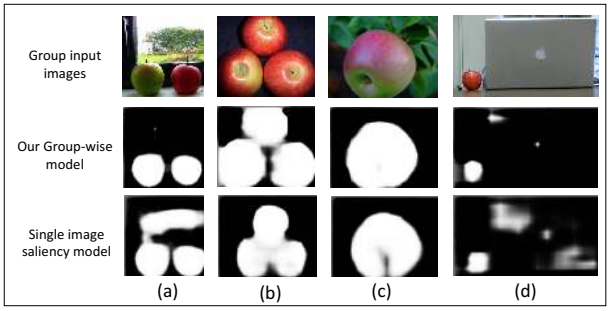

Figure 3: Comparison of our group-wised model and the single image saliency model. With the group-wised feature representation, our model enhanced the common objects and weakened the other parts.

and unique information for the single image. The architecture of the proposed approach is shown in Figure 1.

2.2 Semantic Image Representation

In co-saliency detection, image representation is facing a number of challenges, such as multiple objects, occlusion, and diverse background. More importantly, co-feature representation for image group is the emphasis of our framework. It mainly consists of two component: first, constructing the group-wise feature representation which takes advantage of the intra group theoretical consistency to better obtain the interaction information of group images; and second, computing the single image semantic feature representation for each image individually.

Group-wise Feature Representation

As shown in Figure 1, we adopt a group input and group output FCN to model the group semantic information for a joint representation. The initial high-level semantic feature s_i for each image I_i parameterised by Θ_{shared}:

$$s_i = f_{shared}(I_i; \Theta_{shared}) \tag{3}$$

where f_{shared} is a convolutional process representing shown as the "semantic block" in Figure 1, this block has 13 convolutional layers, it has the parameters Θ_{shared} which are shared among all the semantic blocks. With group input, we generate the shared feature $s = \{s_i\}_{i=1}^{K}$ for each image and these features will be the base on which we will do the next steps, and will be the link between the individual and intra group features since both use it.

Given the image group IG, the problem of group-wise feature representation is converted to the task of how to correspond the related components(such as common objects) defined in a group by their initial feature maps and to learn the interaction between images based on the group consistency.

The next step is concatenating these shared features and then applying 3 convolution layers(shown in Figure 2), this will give the network the possibility to extract the necessary group-wise information that can later be used for the computation of the saliency maps, it is defined as :

$$X = f_{intra}(s; \Theta_{intra}) \tag{4}$$

where Θ_{intra} is the parameters learned from 3 convolutional layers and f_{intra} is the convolutional process representing combining with the concatenation of $s = \{s_i\}_{i=1}^{K}$ shown in the "group saliency block" of Figure 1.

Single Image Feature Representation

The single image features $x = \{x_i\}_{i=1}^{K}$ encode the individual properties for each image I_i. As shown in the "Single feature representation" of Figure 1, taking advantage of the FCN, the feature is generated by a 3 convolutional-layer network. It is defined as follows:

$$x_i = f_{single}(s_i; \Theta_{single}) \tag{5}$$

where Θ_{single} are the parameters learned from the convolutional process f_{single}.

Applying these 3 convolutional layers results in deeper features of each image. These are the features that will be combined with the group-wise features extracted in the previous sub-section. The merging is important because as shown in Figure 3, some objects can be salient, but not present in the entire group, like the tree visible from the window in Figure 3 (a). This shows the necessity of the merging of the two features give the network the necessary flexibility so that it can weaken the saliency map in the regions where a salient object is not present in all the group. The other reason for the merging is enhancing the salient regions for objects that are present in all the image group, this is illustrated by Figure 3 (b) and (c) where the apples which are present in all the images have an increased saliency degree in a group-wise model than in a single image model.

2.3 Collaborative Learning for Image Group

As described previously, we construct the collaborative learning strategy from two components: the group-wise feature learning and the single image individual feature learning, which aims to adaptively capture the interaction relationships between group images and meanwhile retain the characteristics of single image itself. As shown in Figure 1, the collaborative learning structure is discovered through joint learning

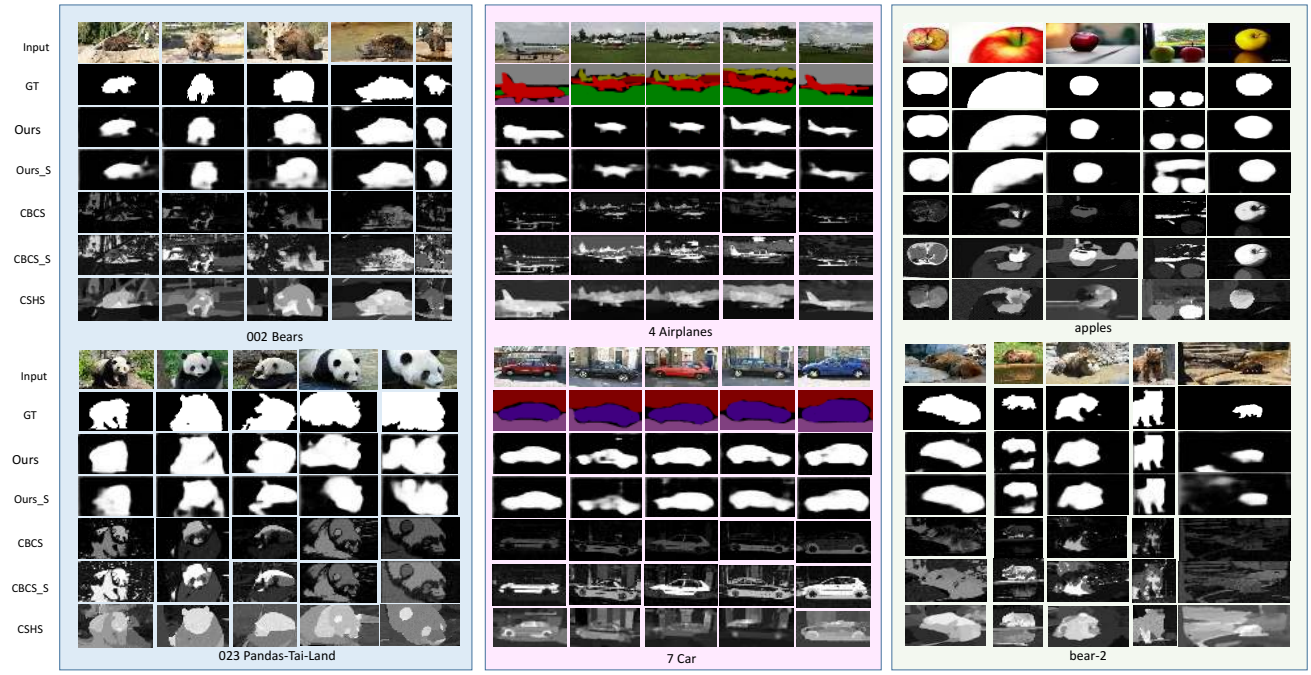

Figure 4: Visual comparison of co-saliency detection on three benchmark datasets. From left to right, the examples are from iCoseg dataset, MSRC dataset, and Cosal2015 dataset, respectively. Obviously, the proposed method performs well in these datsets.

for x and X. Specifically, that means the common object regions are activated by convolutional process and the unique characteristics of single image are weakened but still retained for the final saliency estimation. The merging is defined as:

$$R = f_{collaborative}(x, X; \Theta_{collaborative}) \tag{6}$$

where $f_{collaborative}$ is the function that concatenates each x_i with X, and then applies a convolutional and a deconvolutional layer on each of the results, which gives us the final group saliency, this is illustrated by the "collaborative learning" part of Figure 1, this architecture allows the network to combine the single image features and the group-wise features and obtain the saliency from their combined information.

2.4 Training

In principle, image representation and the learning strategy are correlated and complementary problems which can mutually promote each other. Thus we develop a unified end-to-end data-driven framework with group input and group output, where the group-wise feature and the single image features are learned jointly and adaptively in a supervised setting through the architecture illustrated in Figure 1. For training, all the parameters Θ are learned via minimizing a loss function, which is computed as the errors between the saliency map and the ground truth. Let $\{IG_i\}_{i=1}^N$ and $\{GT_i\}_{i=1}^N$ denote a collection of training samples where N is the number of image groups. Our network is trained by minimizing the following cost function:

$$\sum_{i=1}^{N} \|(GT_i - g(IG_i; \Theta)\|_F^2 \tag{7}$$

where $\Theta = \{\Theta_{shared}, \Theta_{single}, \Theta_{intra}, \Theta_{collaborative}\}$, g is the function that, given an input group, outputs the corresponding saliency maps for it. This cost function corresponds to the squared Euclidean loss term. The network is trained by the stochastic gradient descent (SGD) method to minimize the above cost function, a regularization is applied on all the training samples and all the parameters are learned simultaneously.

3 Experimental Results

3.1 Experimental Setup

Datasets

In order to evaluate the performance of the proposed approach, we conduct a set of qualitative and quantitative experiments on three benchmark datasets annotated with pixel-wised ground-truth labeling, including the iCoseg dataset [Batra *et al.*, 2010], the MSRC-v2 dataset [Winn *et al.*, 2005] and the Cosal2015 dataset [Zhang *et al.*, 2016b]. The iCoseg dataset contains 643 images which divided into 38 groups and they are challenging for co-saliency detection task because of the complex background and multiple co-salient objects. Note that we only use subset5 in this dataset which contains 5 images in each group. Another large dataset widely used in co-saliency detection is the MSRC-v2 datasets which contains 591 images in 23 object classes with man-

ually labeled pixel-wise ground truth data. It is more challenging than iCoseg dataset because of the diverse colors and shapes. The cosal dataset contains 50 image groups and totally 2015 images which are collected from challenging scenarios in the ILSVRC2014 detection benchmark [Russakovsky *et al.*, 2015] and the YouTube video set [Prest *et al.*, 2012].

Implementation Details

The fully convolutional network (FCN) is implemented by using the Caffe [Jia *et al.*, 2014] toolbox. We initialize our network by using a pretrained version of the single image input network (over the MS COCO dataset) which is based on the VGG 16-layer net [Simonyan and Zisserman, 2014] and then, transfer the learned representations by fine-tuning [Donahue *et al.*, 2014] to the co-saliency task by group input and group output. We construct the deconvolution layers by upsampling, whose parameters are initialized as simple bilinear interpolation parameters and iteratively updated during training. We resize all the images and ground-truth maps to 128×256 pixels for training. The momentum parameter is chosen as 0.99, the learning rate is set to 1e-10, and the weight decay is 0.0005. We need about 60000 training iterations for convergence.

The training data we used in our approach are generated from existing image dataset(Coco dataset [Lin *et al.*, 2014]) which has 9213 images with the masks information. In the proposed network, we set up the number of images in each group to 5, namely, $K = 5$. Following the approach of [Siva *et al.*, 2013], we extract Gist and Lab color histogram features, and then calculate the Euclidean distance between images to find 4 other images that are most similar to each one. In this way, we make up training groups. For testing, we randomly sample 5 images from each group as the new image group to ensure the group input size to our model. This sampling procedure proceeds to generate a set of new image groups (with the cardinality being 5) until all the original images are covered in the generated new image groups. For iCoseg dataset, we directly adopt the subset [Batra *et al.*, 2010] which contains 5 images in each group.

3.2 Evaluation Metrics

In the experiments, we utilize four metrics for quantitative performance evaluations, the Precision and Recall (PR) curve, F-measure, mean absolute error (MAE). Specifically, the PR curve reflects the object retrieval performance in precision and recall by binarizing the final saliency map using different thresholds (usually ranging from 0 to 255)[Borji *et al.*, 2015]. The F-measure characterizes the balance degree of object retrieval between precision and recall such that: $F_\eta = \frac{(1+\eta^2)Precision \times Recall}{\eta^2 \times Precision + Recall}$ where η^2 is typically set to 0.3 like the most existing literature work. In addition, MAE refers to the average pixel-wise error between the saliency map and ground truth. Finally, AUC evaluates the object detection performance, and computes the area under the standard ROC curve (false positive rate and true positive rate).

3.3 State-of-the-Art Performance Comparison

In the experiments, we compare the proposed approach with several representative state-of-the-art methods including CSHS [Liu *et al.*, 2014] and CBCS [Fu *et al.*, 2013], whose source codes are publicly available. To investigate the performance differences with and without group interactions, we also make a comparison with the saliency detection approaches for our work and CBCS without group interactions, respectively referred to as Ours_S and CBCS-S. The experimental results are shown in Figure 4. These examples belong to 6 groups of the 3 datasets mentioned above. From the comparison of these examples, we can observe that our proposed approach can better capture the common (in semantic-level) object regions, it also gives more clear borders between the salient and non-salient regions. As shown by the results on the iCoseg image groups which are illustrated on the left set (blue) of Figure 4. The proposed approach does a better job on separating the salient regions and the background with clear boundaries. The middle set (pink) shows the groups in MSRC datast which is mainly for segmentation task. The co-saliency model captures the common objects well, in the semantic level. The right set (green) from Cosal2015 dataset is more challenging that the common objects in this dataset are always in shapes, colors, and viewed from different perspectives. Therefore, our approach performs better than the competing approaches in most cases. Moreover, the proposed group-wise approach with group interactions gives rise to the performance gains relative to the corresponding approach without group interactions.

For quantitative comparison, the PR-curve is shown in Figure 5 on the three datasets, it is observed that our approach performs best in all the datasets. Table 1 shows the comparison between the approaches through different evaluations. In iCoseg dataset and MSRC dataset, the proposed approach performs better than others on most evaluations. In the challenging dataset Cosal2015 (with very complex scene clutters), the proposed approach performs best on all evaluations.

In addition, we make a quantitative performance comparison with some other recently proposed co-saliency approaches over the MSRC dataset, including ESMG, ESMG-S (the variant of ESMG without group interactions), SACS, and CoDW. Since these approaches have no open source codes, we have to directly quote their quantitative results (only having AP scores), which are provided in the work [Zhang *et al.*, 2016b]. As shown in Figure 6, our approach achieves the second best performance in co-saliency detection, and is also comparable to the best CoDW approach (involving many stages and refinement postprocessing operations like manifold rankings). In contrast, our approach is straightforward, end-to-end, and without any postprocessing. Thus, it is a promising choice in practice.

3.4 Analysis of Proposed Approach

As illustrated in Figure 4, our method obtains more robust and complete salient regions. The boundaries of the salient regions are more clear, and in most examples, the proposed approach properly filters the background information. The comparison between our single image model and the group-input group-output model demonstrates the effectiveness and important role of our group-wise feature representation as well as the collaborative learning strategy for the group-wise and single image features, when compared to the single model

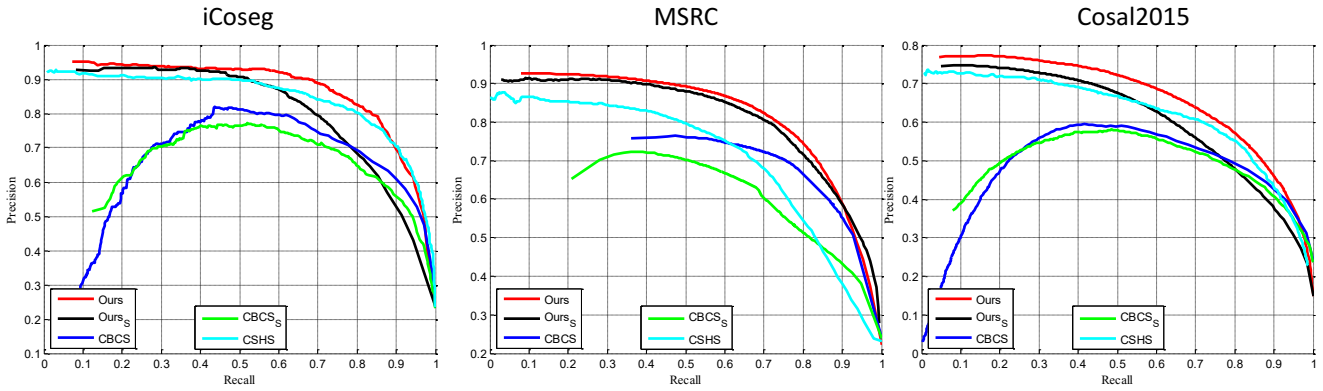

Figure 5: Precision-recall curves of different saliency detection methods on 3 benchmark datasets. Overall, the proposed approach performs well with higher precision in the case of a fixed recall.

Dataset		Ours	Ours_S	CBCS	CBCS_S	CSHS
iCoseg	mF	**0.6983**	0.6935	0.6885	0.6443	0.5288
	AUC	0.9497	0.9256	0.9294	0.9106	**0.9530**
	MAE	**0.1018**	0.1343	0.1922	0.1517	0.1102
MSRC	mF	**0.5952**	0.5671	0.5206	0.5057	0.4612
	AUC	0.6997	0.6799	**0.7030**	0.6981	0.6813
	MAE	**0.2238**	0.2534	0.3677	0.2912	0.2587
Cosal2015	mF	**0.6084**	0.5512	0.5130	0.4942	0.4898
	AUC	**0.8954**	0.8744	0.8261	0.8251	0.8521
	MAE	**0.1434**	0.1611	0.2268	0.1980	0.1883

Table 1: Comparison of mean F-measure using adaptive threshold (mF), AUC scores and MAE scores (smaller better). Our approach achieves the best performance w.r.t. all these metrics in most cases.

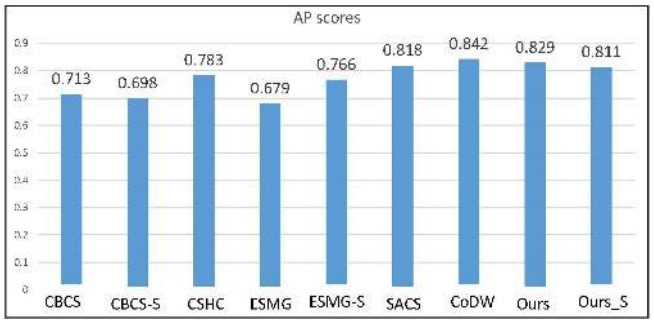

Figure 6: AP scores of different saliency detection methods on MSRC datasets. Overall, the proposed approach performs better in most situations.

approach, the proposed one gives results where the common objects are enhanced and made brighter whereas the different objects are weakened and made dimmer. Meanwhile, we compute the average performance for group-wise model and single image model over all the datasets with respect to F-measure, MAE, and AUC. Overall, our group-wise model respectively achieves 0.6340, 0.8477 and 0.1563 on F-measure, AUC, and MAE, and the single image model correspondingly achieves 0.6039, 0.8266, and 0.1829. This effect is also most clear in Figure 4 on the apples image group where the trees that were detected as salient by the single image model were erased by the group-wise approach because it is not common to all the images of the group.

4 Conclusion

In this paper, we propose a unified deep co-saliency approach for co-salient detection made as a fully convolutional network with group input and group output. It takes a data-driven learning pipeline for capturing the collaboration and consistency intra image group, and subsequently builds an end-to-end learning scheme for explore the intrinsic correlations between the tasks of individual image saliency detection and intra group saliency detection. Through collaborative learning from the co-saliency image group, the deep co-saliency model obtained the capability of capturing the information of both the shared and unique characteristics of each image within the image group and effectively modeled the interaction relationship between them. The experimental results demonstrated that the proposed approach performs favorably in different evaluation metrics against the state-of-the-art methods.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant U1509206 and Grant 61472353, in part by the Alibaba-Zhejiang University Joint Institute of Frontier Technologies.

References

[Batra *et al.*, 2010] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *Proc. IEEE Conf. CVPR*, pages 3169–3176. IEEE, 2010.

[Borji *et al.*, 2015] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Trans. Image Process.*, 24(12):5706–5722, 2015.

[Chang *et al.*, 2011] 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 *Proc. IEEE Conf. CVPR*, pages 2129–2136. IEEE, 2011.

[Chen and Hsu, 2014] Yi-Lei Chen and Chiou-Ting Hsu. Implicit rank-sparsity decomposition: Applications to saliency/co-saliency detection. In *Proc. IEEE Conf. ICPR*, pages 2305–2310. IEEE, 2014.

[Cheng *et al.*, 2014] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu. Salientshape: Group saliency in image collections. *The Visual Computer*, 30(4):443–453, 2014.

[Cho *et al.*, 2015] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *Proc. IEEE Conf. CVPR*, pages 1201–1210, 2015.

[Donahue *et al.*, 2014] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proc. ICML*, volume 32, pages 647–655, 2014.

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

[Fu *et al.*, 2015a] Huazhu Fu, Dong Xu, Stephen Lin, and Jiang Liu. Object-based rgbd image co-segmentation with mutex constraint. In *Proc. IEEE Conf. CVPR*, pages 4428–4436, 2015.

[Fu *et al.*, 2015b] Huazhu Fu, Dong Xu, Bao Zhang, Stephen Lin, and Rabab Kreidieh Ward. Object-based multiple foreground video co-segmentation via multi-state selection graph. *IEEE Transactions on Image Processing*, 24(11):3415–3424, 2015.

[Ge *et al.*, 2016] Chenjie Ge, Keren Fu, Fanghui Liu, Li Bai, and Jie Yang. Co-saliency detection via inter and intra saliency propagation. *Signal Processing: Image Communication*, 44:69–83, 2016.

[Jia *et al.*, 2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proc. ACM Multimedia*, pages 675–678. ACM, 2014.

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

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

[Li *et al.*, 2014] Lina Li, Zhi Liu, Wenbin Zou, Xiang Zhang, and Olivier Le Meur. Co-saliency detection based on region-level fusion and pixel-level refinement. In *Proc. IEEE ICME*, pages 1–6. IEEE, 2014.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. ECCV*, pages 740–755. Springer, 2014.

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

[Prest *et al.*, 2012] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning object class detectors from weakly annotated video. In *Proc. IEEE Conf. CVPR*, pages 3282–3289. IEEE, 2012.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Siva *et al.*, 2013] Parthipan Siva, Chris Russell, Tao Xiang, and Lourdes Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *Proc. IEEE Conf. CVPR*, pages 3238–3245, 2013.

[Tang *et al.*, 2014] Kevin Tang, Armand Joulin, Li-Jia Li, and Li Fei-Fei. Co-localization in real-world images. In *Proc. IEEE Conf. CVPR*, pages 1464–1471, 2014.

[Wang *et al.*, 2015] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *Proc. IEEE Conf. CVPR*, pages 3395–3402, 2015.

[Winn *et al.*, 2005] John Winn, Antonio Criminisi, and Thomas Minka. Object categorization by learned universal visual dictionary. In *Proc. IEEE Conf. ICCV*, volume 2, pages 1800–1807. IEEE, 2005.

[Ye *et al.*, 2015] Linwei Ye, Zhi Liu, Junhao Li, Wan-Lei Zhao, and Liquan Shen. Co-saliency detection via co-salient object discovery and recovery. *IEEE Signal Processing Letters*, 22(11):2073–2077, 2015.

[Zhang *et al.*, 2015] 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 *Proc. IEEE Conf. ICCV*, pages 594–602, 2015.

[Zhang *et al.*, 2016a] Dingwen Zhang, Junwei Han, Jungong Han, and Ling Shao. Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining. *IEEE transactions on Neural Networks and Learning Systems*, 27(6):1163–1176, 2016.

[Zhang *et al.*, 2016b] Dingwen Zhang, Junwei Han, Chao Li, Jingdong Wang, and Xuelong Li. Detection of co-salient objects by looking deep and wide. *Int. J. Comput. Vis.*, 120(2):215–232, 2016.