ADAPTIVE INTRA-GROUP AGGREGATION FOR CO-SALIENCY DETECTION

Guangyu Ren [*], *Tianhong Dai* [*], *and Tania Stathaki*

Imperial College London, United Kingdom

ABSTRACT

Co-salient object detection (CoSOD) together with the rapid development of deep learning has led to substantial progress in recent years. However, the feature aggregation between group feature representation and individual feature representation is still a challenging issue. In this work, we propose a novel adaptive intra-group aggregation (AIGA) method, which provides a new perspective to investigate the interaction relationship between group and single-image features and aggregate these features in an adaptive way. A novel scale-aware loss is proposed to help the model capture the scale prior of different groups and discriminatively process groups during the training phase. Extensive experiments demonstrate that the proposed method can effectively improve the performance without increasing extra parameters and achieve better accuracy on three prevalent benchmarks.

Index Terms— Co-salient object detection, determinantal point processes

1. INTRODUCTION

Salient object detection (SOD) aims at imitating the human vision system to segment prominent objects in a given image. As a new attractive branch of the SOD tasks, co-salient object detection (CoSOD) aims at detecting common foreground and co-occurring objects in a group of images. Due to the potential use of detecting objects in multiple relevant images, CoSOD can be treated as a pre-processing step and has been widely employed in diverse computer vision tasks, such as image retrieval [1], image surveillance [2], co-segmentation [3] and weakly-supervised segmentation [4]. One of the key challenging issues of this task is modeling the relationship between individual feature representation in a single image and group feature representation. Wei [5] investigates the interaction relationship between single image and group images by processing them separately in different network blocks to obtain the unique characteristics of each single image and group consistency. Zhang [6] extracts the intra- and inter- image correspondence of an image group by designing an adaptive graph convolutional network. Both these methods model the interaction relationship in suboptimal ways due to the requirement of extra parameters and networks. This issue motivates us to propose a simple but effective algorithm which can formulate the discrepancy between group-wise and individual feature representations and meanwhile adaptively fuse these features. In addition, due the size variation among different salient objects, we propose a scale-aware loss to further refine the detection performance.

2. RELATED WORKS

2.1. Co-Salient Object Detection

Accurate detection performance has been achieved on deep learning methods, which jointly learn co-salient object representations [7]. Zha [8] proposes a new end-to-end approach to learn single-image features with pyramid spatial attention operation. Subsequently, a hierarchical low-rank bilinear pooling function is employed to learn group semantic representation for co-category classification. Li [9] designs a co-attention recurrent unit to model the final group representation, which is treated as synergetic information and fused with each individual image.

2.2. Determinantal Point Processes

In this work, determinantal point processes (DPPs) are used to model the diversity within intra-group. Recently, DPPs have been widely used in the machine learning community, such as video summarization [10], recommendation systems [11], and deep reinforcement learning [12, 13].

In general, for a discrete set of points $\mathcal{Y} = \{x_1, x_2, \cdots, x_N\}$, a point process $\mathcal{P}$ is a probability measure over all $2^{|\mathcal{Y}|}$ subsets. $\mathcal{P}$ is a DPP if a random subset $\mathbf{Y}$ is sampled with probability:

$$\mathcal{P}_L(\mathbf{Y} = Y) = \frac{\det(L_Y)}{\sum_{Y' \subseteq \mathcal{Y}} \det(L_{Y'})} = \frac{\det(L_Y)}{\det(L+I)}, \qquad (1)$$

where $Y \subseteq \mathcal{Y}$, I is the identity matrix, $L \in \mathbb{R}^{N \times N}$ is the positive semi-definite DPP kernel matrix, and L_Y is the submatrix with rows and columns indexed by the elements of the subset Y.

The kernel matrix L can be represented as the Gram matrix $L = X^T X$, where each column of X is the feature vector of an item in $\mathcal{Y}$. $\det(L_Y)$ therefore represents the (squared)

[*] Equal Contributions

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

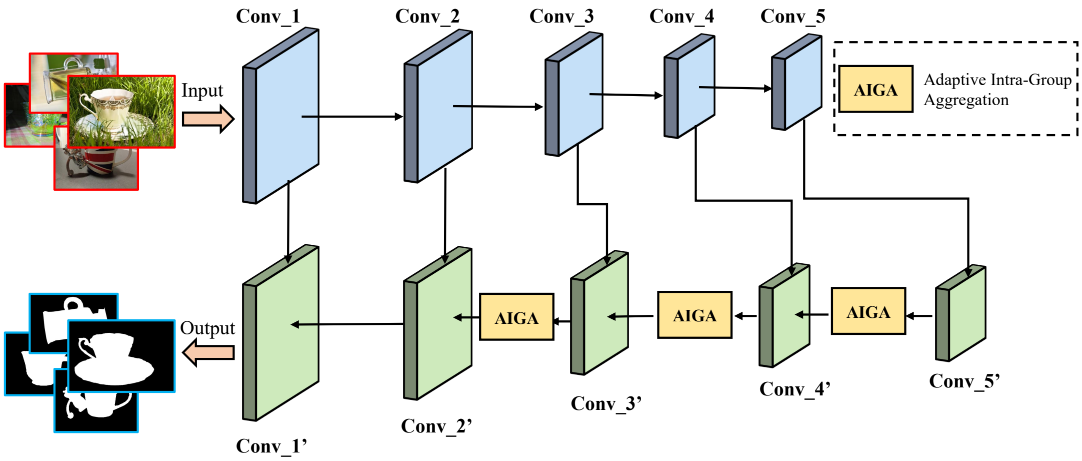

Fig. 1: Overall structure of the proposed method.

volume spanned by vectors $x_i \in Y$. From a geometrical perspective, feature vectors in a subset that are more orthogonal will have a larger determinant, increasing their probability of being sampled in a DPP: $\mathcal{P}_L(\mathbf{Y} = Y) \propto \det(L_Y)$. Using orthogonality as a measure of diversity, we adopt DPPs to model the diversity of the samples in the intra-group.

3. METHODOLOGY

The overall architecture is shown in Figure 1. We simply adopt the Feature Pyramid Network (FPN) [14] structure with a VGG-16 [15] backbone as our final network. Considering the computing resources, the proposed AIGAs are only applied on deep layers, namely, $Conv5^{'}$, $Conv4^{'}$ and $Conv3^{'}$.

3.1. Adaptive Intra-Group Aggregation

We are given a group of images $G = \{I_n\}_{n=1}^{N}$ as the input for the encoder network $F(\cdot)$, such as a VGG-16 backbone. This network extract features from group images and the feature representation of each image can be obtained:

$$e_n = F(I_n) \tag{2}$$

Then we simply concatenate individual feature representations together and adopt a mean operation to effectively obtain the group feature representation:

$$E = \frac{1}{N}\sum_{n=1}^{N} e_n \tag{3}$$

where E indicates the group feature representation of N images. Then we calculate the diversity between the group G and each subset I_n by DPP:

$$\alpha = DPP(e_n, E) \tag{4}$$

Subsequently, we treat the diversity α of e_n and E as the discrepancy between group features and individual features and adaptively fuse group features with each image according to the discrepancy:

$$\hat{e_n} = e_n + \alpha \cdot E \tag{5}$$

where $\hat{e_n}$ indicates the final individual feature representation. With the proposed AIGA, the processed individual features not only contain unique characteristics of single image but the common foreground information of group images.

3.2. Scale-aware Loss

To better capture the scale variation of salient objects among different groups, we design a scale-aware loss by calculating the scale prior of each group. Concretely, the scale prior can be obtained as follows:

$$S = \frac{1}{N}\sum_{n=1}^{N} GT_n \tag{6}$$

where GT_n refers to the corresponding ground truth of I_n. Therefore, S indicates the scale prior of a specific group with N images. Due to the imbalance of objects' sizes, the model should treat each group unequally in the training phase. More specifically, a large punishment should be assigned on training loss when S is small and vice versa. The adaptive punishment weight of each group is formulated as follow:

$$\beta = (1 - S)^p \tag{7}$$

Here we use $p = 0.1$ to control the variation of β. Finally, the training loss can be expressed as:

$$L_{iou}(GT_n, P) = \beta \cdot L_{iou}(GT_n, P) \tag{8}$$

where P is the prediction of each image in group G, L_{iou} refers to IOU loss function. To this end, the model is able

Table 1: Quantitative comparisons through the maximum of F-score F_β, S-score S_α, E-score E_θ, and error-score M, over three widely evaluated datasets.

	Metric	CBCS [16]	GWD [5]	RCAN [17]	ESMG [18]	BASNet [19]	UMLF [20]	CSMG [21]	EGNet [22]	SCRN [23]	Ours *
CoSOD3k	$F_\beta \uparrow$	0.466	0.649	0.688	0.418	0.720	0.639	0.709	0.702	0.716	0.746
	$S_\alpha \uparrow$	0.528	0.716	0.744	0.523	0.771	0.632	0.711	0.762	0.771	0.788
	$E_\theta \uparrow$	0.637	0.777	0.808	0.635	0.804	0.758	0.804	0.793	0.805	0.836
	$M \downarrow$	0.228	0.147	0.130	0.239	0.114	0.285	0.157	0.119	0.113	0.088
Cosal2015	$F_\beta \uparrow$	0.523	0.706	0.764	0.476	0.791	0.690	0.784	0.786	0.783	0.812
	$S_\alpha \uparrow$	0.544	0.744	0.779	0.522	0.822	0.662	0.774	0.818	0.817	0.829
	$E_\theta \uparrow$	0.656	0.802	0.842	0.640	0.849	0.769	0.842	0.843	0.850	0.865
	$M \downarrow$	0.233	0.148	0.126	0.247	0.096	0.271	0.130	0.099	0.098	0.080
CoCA	$F_\beta \uparrow$	0.313	0.408	0.422	-	0.408	-	0.499	0.404	0.413	0.466
	$S_\alpha \uparrow$	0.523	0.602	0.616	-	0.592	-	0.627	0.603	0.612	0.628
	$E_\theta \uparrow$	0.641	0.701	0.702	-	0.644	-	0.733	0.648	0.642	0.681
	$M \downarrow$	0.180	0.166	0.160	-	0.195	-	0.114	0.178	0.164	0.148

to effectively capture the relative scale information of salient objects in a specific group.

4. EXPERIMENTS

4.1. Datasets and Evaluation Metrics

We adopt three current challenging datasets for evaluating the effectiveness of the proposed algorithms: CoCA [24], CoSOD3k [7] and Cosal2015 [25]. In order to comprehensively evaluate the detection results, we use four prominent evaluation metrics, namely, the F-measure (F_β), Mean Absolute Error (M), E-measure (E_θ) and S-measure (S_α). More concretely, F_β is computed by pairs of precision and recall on different thresholds from 0 to 255:

$$F_\beta = \frac{\left(1+\beta^2\right) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \tag{9}$$

where β^2 is set to 0.3 as default. M indicates the pixel-wise absolute error between the predicted saliency map P and the corresponding ground truth G:

$$M = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \left| P\left(x,y\right) - G\left(x,y\right) \right| \tag{10}$$

where H and W denote the height and width of the saliency map respectively. In addition, S_α represents the structural similarity and E_θ captures image-level statistics and their local pixel matching information.

4.2. Implementation Details

We adopt FPN with a VGG-16 backbone as our baseline. We use DUTS[26] dataset in the training phase and follow the same group setting in [24]. The final model is trained on a GTX TITAN X GPU for 50 epochs with the Adam optimizer.

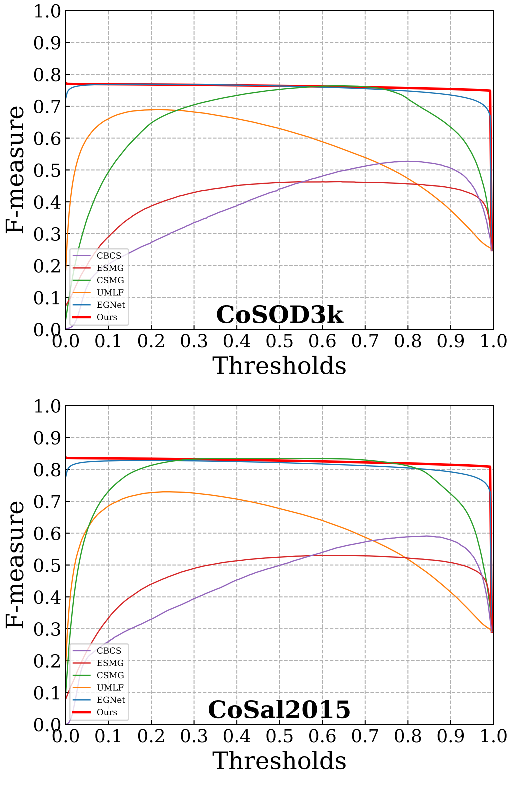

Fig. 2: F-measure curves on two public CoSOD datasets.

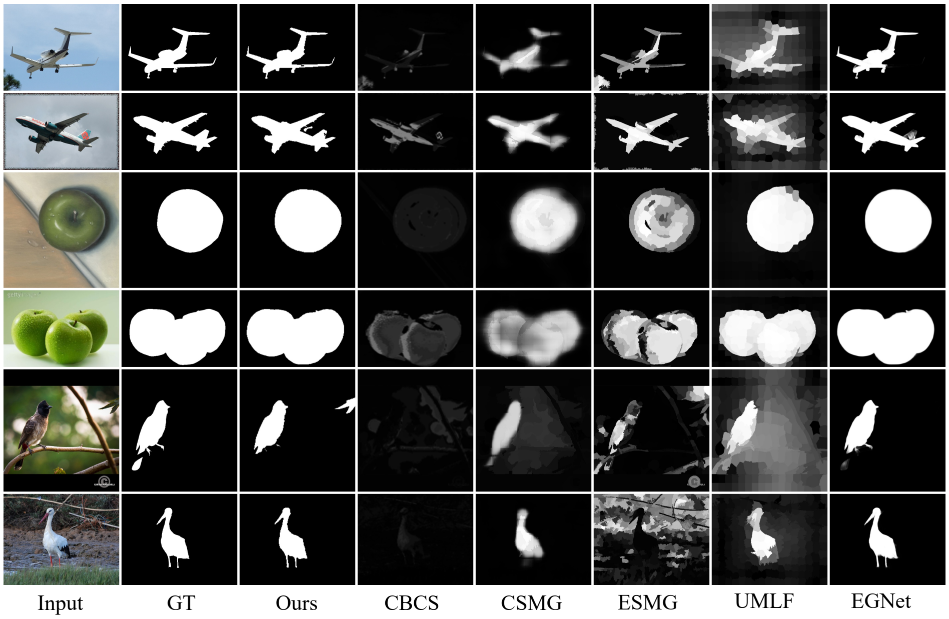

Fig. 3: Qualitative comparisons with state-of-the-arts in different categories.

Table 2: Ablation studies on the proposed methods.

	Metric	Base	AIGA5	AIGA54	AIGA543	+Scale
CoSOD3k	$F_\beta \uparrow$	0.731	0.742	0.743	0.742	0.746
	$S_\alpha \uparrow$	0.774	0.780	0.781	0.780	0.788
	$E_\theta \uparrow$	0.825	0.832	0.833	0.834	0.836
	$M \downarrow$	0.096	0.088	0.089	0.089	0.088
Cosal2015	$F_\beta \uparrow$	0.798	0.807	0.812	0.815	0.812
	$S_\alpha \uparrow$	0.817	0.820	0.823	0.825	0.829
	$E_\theta \uparrow$	0.856	0.861	0.864	0.867	0.865
	$M \downarrow$	0.088	0.086	0.084	0.082	0.080

The initial learning rate is set to $1e-4$ and the batch size is 16. All images are reshaped to 224x224 in the training and testing stages and reshape to the original size in the evaluation stage.

4.3. Performance Comparison

We compare our method with 9 challenging state-of-the-art methods, including representative traditional models and deep learning models, namely, CBCS [16], ESMG [18], UMLF [20], GWD [5], RCAN [17], CSMG [21], EGNet [22], SCRN [23] and BASNet [19]. We directly use the public saliency maps and evaluation results for fair comparisons.

Quantitative Evaluation. We show F-measure curves in Figure 2 to evaluate the overall performance of the CoSOD. As shown, our method which is represented by the red line outperforms other state-of-the-arts. In Table 1, compared with previous approaches, our algorithm achieves competitive results across all datasets, especially on CoSOD3k and Cosal2015, which show the best results on four evaluation metrics.

Qualitative Evaluation. Figure 3 shows the visual results of the proposed method and other representative methods. It can be observed that our saliency maps are more similar to the ground truth and show better performance on completeness and edges in different groups.

Ablation Analysis. We conduct experiments to show the effectiveness of each proposed algorithm. As shown in Table 2, $Base$ refers to the FPN baseline. We apply the AIGA on different levels from deep to shallow layers. For instance, $AIGA5$ indicates we only apply the proposed AIGA on the deepest level $Conv5^{'}$. It can be clearly observed that the proposed AIGA can effectively improve the detection results and three levels ($AIGA543$) show the better results. $Scale$ represents that the scale-aware loss is applied on the model, leading to further improvement of the detection results. Apparently, it is demonstrated that both proposed methods can effectively improve the accuracy without incorporating additional parameters.

5. CONCLUSION

In this paper, we first propose a novel feature aggregation algorithm for co-saliency detection. Our adaptive intra-group aggregation utilizes determinantal point process to formulate the discrepancy between group features and individual features and simply combine these features in an adaptive way. We also propose a scale-aware loss to capture the scale variation and provide scale prior of each group to the model. Extensive experiments demonstrate that both methods can effectively boost the detection performance.

6. REFERENCES

[1] Guanghai Liu and Dengping Fan, "A model of visual attention for natural image retrieval," in *ISCC-C*, 2013.

[2] Zhifan Gao, Chenchu Xu, Heye Zhang, Shuo Li, and Victor Hugo C de Albuquerque, "Trustful internet of surveillance things based on deeply represented visual co-saliency detection," *IEEE IoT-J*, 2020.

[3] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang, "Deepco3: Deep instance co-segmentation by co-peak search and co-saliency detection," in *CVPR*, 2019.

[4] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang, "Joint learning of saliency detection and weakly supervised semantic segmentation," in *ICCV*, 2019.

[5] Lina Wei, Shanshan Zhao, Omar El Farouk Bourahla, Xi Li, and Fei Wu, "Group-wise deep co-saliency detection," *arXiv preprint arXiv:1707.07381*, 2017.

[6] Kaihua Zhang, Tengpeng Li, Shiwen Shen, Bo Liu, Jin Chen, and Qingshan Liu, "Adaptive graph convolutional network with attention graph clustering for co-saliency detection," in *CVPR*, 2020.

[7] Deng-Ping Fan, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Huazhu Fu, and Ming-Ming Cheng, "Taking a deeper look at co-salient object detection," in *CVPR*, 2020.

[8] Zheng-Jun Zha, Chong Wang, Dong Liu, Hongtao Xie, and Yongdong Zhang, "Robust deep co-saliency detection with group semantic and pyramid attention," *IEEE TNNLS*, 2020.

[9] Bo Li, Zhengxing Sun, Qian Li, Yunjie Wu, and Anqi Hu, "Group-wise deep object co-segmentation with co-attention recurrent neural network," in *ICCV*, 2019.

[10] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *CVPR*, 2017.

[11] Yichao Wang, Xiangyu Zhang, Zhirong Liu, Zhenhua Dong, Xinhua Feng, Ruiming Tang, and Xiuqiang He, "Personalized re-ranking for improving diversity in live recommender systems," *arXiv preprint arXiv:2004.06390*, 2020.

[12] Tianhong Dai, Hengyan Liu, Kai Arulkumaran, Guangyu Ren, and Anil Anthony Bharath, "Diversity-based trajectory and goal selection with hindsight experience replay," in *PRICAI*, 2021.

[13] Tianhong Dai, Yali Du, Meng Fang, and Anil Anthony Bharath, "Diversity-augmented intrinsic motivation for deep reinforcement learning," *Neurocomputing*, 2022.

[14] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.

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

[16] Huazhu Fu, Xiaochun Cao, and Zhuowen Tu, "Cluster-based co-saliency detection," *IEEE TIP*, 2013.

[17] Bo Li, Zhengxing Sun, Lv Tang, Yunhan Sun, and Jinlong Shi, "Detecting robust co-saliency with recurrent co-attention neural network.," in *IJCAI*, 2019.

[18] Yijun Li, Keren Fu, Zhi Liu, and Jie Yang, "Efficient saliency-model-guided visual co-saliency detection," *IEEE SPL*, 2014.

[19] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand, "Basnet: Boundary-aware salient object detection," in *CVPR*, 2019.

[20] Junwei Han, Gong Cheng, Zhenpeng Li, and Dingwen Zhang, "A unified metric learning-based framework for co-saliency detection," *IEEE TCSVT*, 2017.

[21] Kaihua Zhang, Tengpeng Li, Bo Liu, and Qingshan Liu, "Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing," in *CVPR*, 2019.

[22] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng, "Egnet: Edge guidance network for salient object detection," in *ICCV*, 2019.

[23] Zhe Wu, Li Su, and Qingming Huang, "Stacked cross refinement network for edge-aware salient object detection," in *ICCV*, 2019.

[24] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng, "Gradient-induced co-saliency detection," in *ECCV*, 2020.

[25] Dingwen Zhang, Junwei Han, Chao Li, Jingdong Wang, and Xuelong Li, "Detection of co-salient objects by looking deep and wide," *IJCV*, 2016.

[26] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan, "Learning to detect salient objects with image-level supervision," in *CVPR*, 2017.