Detecting Robust Co-Saliency with Recurrent Co-Attention Neural Network

Bo Li[1] , **Zhengxing Sun**[1*] , **Lv Tang**[1] , **Yunhan Sun**[2] and **Jinlong Shi**[2*]
[1]State Key Lab for Novel Software Technology, Nanjing University, Nanjing, China
[2]Jiangsu University of Science and Technology Zhenjiang, China
njumagiclibo@gmail.com, szx@nju.edu.cn, tanglv@smail.nju.edu.cn, 347017917@qq.com,
jlshifudan@gmail.com

Abstract

Effective feature representations which should not only express the image's individual properties, but also reflect the interaction among group images are essentially crucial for robust co-saliency detection. This paper proposes a novel deep learning co-saliency detection approach which simultaneously learns single image properties and robust group feature in a recurrent manner. Specifically, our network first extracts the semantic features of each image. Then, a specially designed Recurrent Co-Attention Unit (RCAU) will explore all images in the group recurrently to generate the final group representation using the co-attention between images, and meanwhile suppresses noisy information. The group feature which contains complementary synergetic information is later merged with the single image features which express the unique properties to infer robust co-saliency. We also propose a novel co-perceptual loss to make full use of interactive relationships of whole images in the training group as the supervision in our end-to-end training process. Extensive experimental results demonstrate the superiority of our approach in comparison with the state-of-the-art methods.

1 Introduction

In practice, the frequently occurring patterns or the prime foregrounds can be utilized to represent the main content of the image group. The co-saliency task is derived with the goal of discovering the common and salient objects in a group of related images. Unlike traditional saliency detection [Li *et al.*, 2019; Hou *et al.*, 2019; Wang *et al.*, 2018; Zhang *et al.*, 2017b], co-saliency explore more synergetic information from multiple images. It endows many high-level computer vision systems with the capability to fixate their attention on the most valuable information in the image group, such as image co-segmentation [Chang *et al.*, 2011], object co-localization [Xue *et al.*, 2013], and video foreground extraction [Fu *et al.*, 2015].

As a fundamental but challenging research topic, to detect co-saliency accurately, there are two key issues must be concerned: i) extract and learn effective feature representations of images in the group; ii) model the synergetic relationship among the co-salient regions at the group level to generate the final co-saliency maps. Conventional approaches [Fu *et al.*, 2013; Liu *et al.*, 2014] utilize handcrafted features, such as color, texture and SIFT descriptors etc., and these methods rely on researcher's prior knowledge to model the interaction between the group images, like inter-image saliency. However, low-level features and fixed hand-designed interaction models are too subjective to face the multiple challenges including background clutter, appearance variance of co-salient object across images, and similarity between co-object and non-common object, etc. Deep learning has recently emerged and demonstrated success in many computer vision applications. Recent researches use deep visual features to improve co-saliency detection and they also try to learn more robust co-salient properties among images in a data driven manner. However, as feature extraction and co-saliency detection are treated as separate steps in these approaches [Zhang *et al.*, 2016; Zhang *et al.*, 2017a; Li *et al.*, 2017; Zheng *et al.*, 2018], they can't customize features for better inferring co-salient regions, leading to suboptimal performance. For robust co-saliency detection, the feature representation should not only express the individual properties of each image, but also contain the relevance and interaction among group images. Recently, [Wei *et al.*, 2017] integrated feature learning and detection as a whole process and proposed an end-to-end deep learning co-saliency detection method. The designed group-wise representation can capture the interaction among group images and improve the results, which demonstrates its importance for co-saliency detection.

As the first end-to-end co-saliency network, [Wei *et al.*, 2017] had made a remarkable success, however in real-world robust co-saliency detection this kind of network architecture suffers from several drawbacks. i) Their end-to-end network architecture requires constant input data, so they construct the network based on the assumption that there are a constant number of images in each group, which are set to be 5 in their work. However, the size of each group is not fixed in real-world scenarios as well as the experimental co-saliency dataset. When encountering the group with various

[*] Correspondence should be addressed to Zhengxing Sun and Jinlong Shi

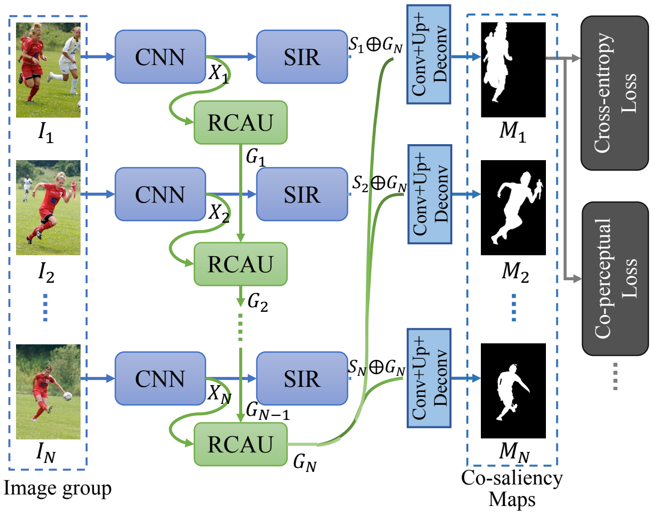

Figure 1: Illustration of the proposed recurrent network architecture for co-saliency detection. ⊕ means concatenate operation.

sizes, they either can't make the detection or only can capture incomplete synergetic relationships with partial group-wise features, limiting the robustness of the model. ii) As there is unpredictable appearance and location variance of the co-salient object across images, their group-wise representation which is based on the direct concatenation of individual image features contains much noise information of un-salient or non-common regions, leading to unsatisfying results.

In this paper, in order to detect robust co-saliency, we propose a novel deep learning approach with recurrent co-attention network. Our aim is to make use of all available information including individual image properties as well as the group synergetic information to create a robust and effective co-saliency network. Specifically, our network will first extract the semantic features of all images, then will be processed by two branches. The single image representation branch processes each image individually to learn the unique properties. Since there are an unfixed number of images in each group and the appearance as well as the location of co-salient object varies across images, we specially design a novel Recurrent Co-Attention Unit (RCAU) to learn the group information. By using the spatial and channel co-attention between images, the RCAU recurrently explores all images in the group to gradually learn the robust group representation and meanwhile suppresses noisy information. The group feature is then broadcasted to each individual image, which allows the network to leverage the synergetic information and unique properties between the images. So the complementarity and interaction of group and single representation are sufficiently exploited to facilitate the robust co-saliency reasoning. Moreover, to make full use of the interactive relationships of whole images in the training group, inspired by [Hsu *et al.*, 2018], we further propose a novel co-perceptual loss as the additional supervision in our end-to-end training process. As far as we know, this is the first attempt to introduce the recurrent architecture into deep co-saliency detection, which allows the proposed method to make full use of all images information in various size groups without retraining the model or introducing extra parameters. Extensive experimental results demonstrate the superiority of our approach in robust co-saliency detection as compared to the state-of-the-art methods.

2 Proposed Approach

2.1 Problem Formulation

Co-saliency detection aims at discovering the common and salient objects in a group of N relevant images $\mathcal{I} = \{I_n\}_{n=1}^N$. The co-saliency maps $\mathcal{M} = \{M_n\}_{n=1}^N$ are produced by a co-saliency detection model:

$$\mathcal{M} = F(\mathcal{I}; \Theta), \tag{1}$$

where $F()$ is the model function that takes an image group as input and outputs a group of co-saliency maps simultaneously. Θ represents model parameters which are optimized by an end-to-end learning scheme in this work. The core idea of this work is trying to make full use of all available information to learn the effective feature representations which can not only express the image's individual properties, but also reflect the interaction among group images for robust co-saliency referring. The overall architecture of the proposed approach is illustrated in Figure 1. For an input image group with an arbitrary size, our network will first extract the semantic features of all images. Then the single image representation (SIR) branch processes each image individually to learn the unique properties. Meanwhile the Recurrent Co-Attention Unit (RCAU) recurrently explores all images in the group to learn the robust group representation. The two branches are later merged for the final co-saliency detection.

2.2 Single Image Representation

As a basic rule in co-saliency, in most cases, the co-salient regions should be salient with respect to the background in each image. So, it is important to learn the unique properties of each image to capture the saliency of the potential co-salient objects in the individual image. In the proposed approach, for each image I_n in the input group $\mathcal{I}$ we first use a pre-trained convolutional neural network (CNN) to extract its semantic features $X_n \in \mathcal{R}^{H \times W \times C}$. Then we construct an SIR block with 3 convolutional layers to encode the individual properties for each image $S = \{S_n\}_{n=1}^N$, which is defined as follows:

$$S_n = f_S(X_n; \Theta_S), \tag{2}$$

where Θ_S are the parameters learned from the convolutional process f_S.

2.3 Group Representation with RCAU

As images within a co-saliency group are contextually associated with each other in different ways such as common objects, similar categories, and related scenes, learning a robust group representation which contains the relevance and interaction between group images is extremely important for co-saliency referring. In this work, we proposed to use a recurrent architecture to learn the group representation G_N for an arbitrary size group $\mathcal{I}$. It is defined as follows:

$$G_N = f_G(\{X_n\}_{n=1}^N; \Theta_G) \tag{3}$$

where Θ_G are the parameters learned from the recurrent convolutional process f_G. Since there is much noise information of non-salient or non-common regions in the group and the appearance as well as the location of co-salient object varies across images, we specifically design a novel recurrent unit RCAU to gradually explore all the synergetic relationships between images for the group representation. As illustrated in Fig. 2(a), for step n, RCAU takes two inputs: the current image feature map X_n and the group representation G_{n-1} of all the images just been explored. In the first step, G_0 is initialized with X_1. We construct two gates in our RCAU, the reset gate g_d is used for denoising current image and the update gate g_z is used to decide how to update current group representation G_n.

Since all images in the group share the common objects, we want to use the synergetic relationships between the explored images and the current image to suppress the noise data in current image, like the non-salient background and non-common regions. The reset gate D is defined as:

$$D = g_d([G_{n-1}, X_n]). \tag{4}$$

Then the denoised feature map $\tilde{X}_n$ is computed as

$$\tilde{X}_n = X_n \odot D = X_n \odot sigmoid(\mathbf{W}_d \times [G_{n-1}, X_n]), \tag{5}$$

where $[G_{n-1}, X_n] \in \mathcal{R}^{H \times W \times 2C}$ is the concatenated feature map of G_{n-1} and X_n, $\times$ is matrix multiplication and $\odot$ denotes element-wise multiplication. We use a FC layer $\mathbf{W}_d$ to reduce the feature dimension.

As the appearance and the location of co-salient object varies across images, we want to fully explore the spatial-channel-wise variation of the co-salient object with co-attention mechanism to determine what group information should be retained in G_{n-1} and what new information should be updated from $\tilde{X}_n$. So $Z_n = (G_{n-1} - X_n\ )$ is used as an input of g_z to model the cross-images variation in each step. The update gate Z is defined as follows,

$$Z = g_z(Z_n; \Theta_z) = g_z(G_{n-1} - X_n). \tag{6}$$

Fig.2(b) illustrates the structure of update gate model g_z.

For spatial attention model, we first use a global cross-channel average pooling layer to get the overall response in each spatial position.Then two FC layers (the first layer is followed by ReLU) are applied to generate the spatial attention maps $Z_s \in \mathcal{R}^{H \times W \times 1}$. It is formulated as

$$Z_s = \mathbf{W}_s^2 \times ReLU(\mathbf{W}_s^1 \times Z_n^{H,W}) \tag{7}$$

where $Z_n^{H,W} \in \mathcal{R}^{H,W}$ is the result of Z_n after cross-channel average pooling.

For channel attention model, we first introduce a global spatial space average pooling layer to get overall response of each channel. Then a FC layer is applied to get the channel attention maps $Z_c \in \mathcal{R}^{1 \times 1 \times C}$, which is formulated as

$$Z_c = \mathbf{W}_c \times Z_n^C, \tag{8}$$

where $Z_n^C \in \mathcal{R}^C$ is the result of Z_n after global spatial space average pooling.

The overall attention maps of current input feature are the product of spatial attention maps and channel attention maps.

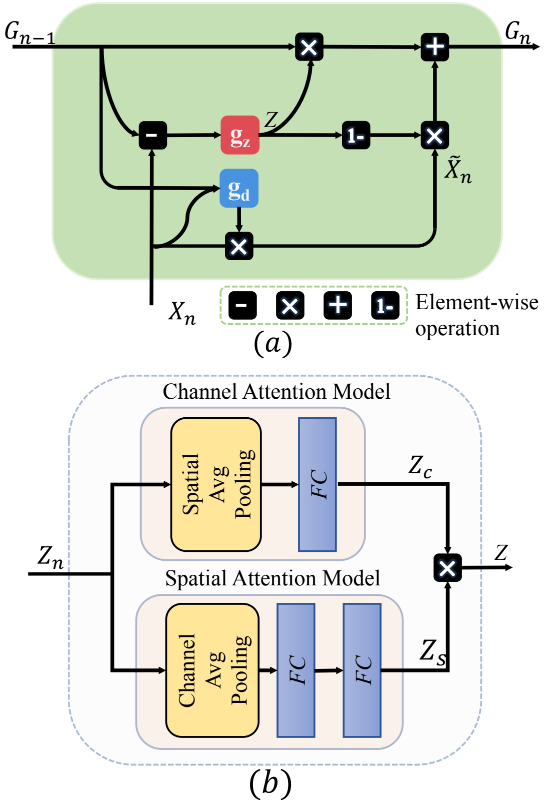

Figure 2: (a) The architecture of recurrent co-attention unit (RCAU). (b) The architecture of proposed update gate model g_z.

After a $sigmoid$ operation, the overall attention maps are normalized into the range between 0 and 1, formulated as

$$Z = sigmoid(\ Z_s \odot Z_c\ ), \tag{9}$$

where $Z \in \mathcal{R}^{H \times W \times C}$. Then the RCAU updates G_n by

$$G_n = Z \odot G_{n-1} + (1 - Z) \odot \tilde{X}_n. \tag{10}$$

The value of each position in Z denotes the probability for the activation value in corresponding position of G_{n-1} to be reserved and position of $\tilde{X}_n$ to be updated. Higher probability value indicates that the update gate model g_z considers that last group feature G_{n-1} in this location has high quality group information and should be reserved. While locations with lower probability means there is new useful synergetic information in $\tilde{X}_n$ should be updated to G_n. Unlike the conventional recurrent units (GRU or LSTM), our RCAU can use co-attention to recurrently learn robust group representation for co-saliency referring meanwhile reducing the noise data. We will provide justification on this issue in the later experiments section.

2.4 Detect Co-saliency with Robust Representation

As described previously, the group feature is then broadcasted to each individual image, which allows the network to leverage the synergetic information and unique properties between the images. So the complementarity and interaction of group and single representation are sufficiently exploited to facilitate the robust co-saliency reasoning. We concatenate each S_i with G_N to get final fused features, and to alleviate the aliasing effect of upsampling, we add a 3×3 convolutional layer after merging operations. So the co-saliency maps $\mathcal{M}$ can be obtained with the final fused features by applying 3×3 convolutional layers and deconvolutional layers followed with a sigmoid activation function. It is formulated as:

$$\mathcal{M} = f_C(S, G_N; \Theta_C). \tag{11}$$

3 Loss Function

Let $\mathcal{I} = \{I_n\}_{n=1}^N$ and their groundtruth $\{\mathcal{G}_n^T\}_{n=1}^N$ denote a collection of training samples where N is the number of images. After co-saliency detection, co-saliency maps are $\{M_n\}_{n=1}^N$. We use the *sigmoid* cross-entropy loss as the individual supervision for each image I_n:

$$L_s(I_n; \Theta) = -\big(\, \mathcal{G}_n^T log(M_n) + (1 - \mathcal{G}_n^T) log(1 - M_n) \,\big). \tag{12}$$

To fully explore the group information, inspired by the wildly used perceptual loss in Neural Style Transfer (NST) [Gatys *et al.*, 2016] works, we proposed a novel co-perceptual loss to model the synergetic relationships between the co-salient objects in training images as an additional supervision. As defined in NST, the core idea of feature perceptual loss is to seek the consistency of two images between the hidden representations in a pre-trained CNN ϕ [Simonyan and Zisserman, 2014]. In this work, for a image I_n, we can generate two masked images with its co-saliency map M_n and groudtruth $\mathcal{G}_n^T$

$$I_n^o = M_n \otimes I_n \quad and \quad I_n^{\mathcal{G}} = \mathcal{G}_n^T \otimes I_n, \tag{13}$$

where $\otimes$ denotes element-wise multiplication. The masked image I_n^o means our detected co-salient regions of I_n while image $I_n^{\mathcal{G}}$ means the real co-salient regions of I_n. Then we apply the extractor ϕ to all masked images $\{I_n^o, I_n^{\mathcal{G}}\}_{n=1}^N$ and obtain their corresponding hidden representations $\{\phi(I_n^o) \in \mathcal{R}^V, \phi(I_n^{\mathcal{G}}) \in \mathcal{R}^V\}_{n=1}^N$, where V is the feature dimension. So, in the training process, we construct the co-perceptual loss for image I_n by measuring the similarity between its detected region I_n^o and the real co-salient regions of all the rest images in the group $I_m^{\mathcal{G}}$ ($I_m \in \mathcal{I}$ and $m \neq n$). The co-perceptual loss function corresponds to the squared Euclidean loss term and is defined by

$$L_c(I_n, I_{m \neq n}) = \frac{1}{V(N-1)} \sum_{m \neq n} || \, \phi(I_n^o) - \phi(I_m^{\mathcal{G}}) \, ||_F^2 \tag{14}$$

Note that all parts of our network are trained jointly, and the over all loss function is given as

$$L = \sum_{n=1}^{N} \big(\, L_s(I_n; \Theta) + \lambda \cdot L_c(I_n, I_{m \neq n}; \Theta) \,\big), \tag{15}$$

where λ is the tradeoff parameter and $\Theta = \{\Theta_S, \Theta_G, \Theta_C\}$ is the all learnable parameters set of our network.

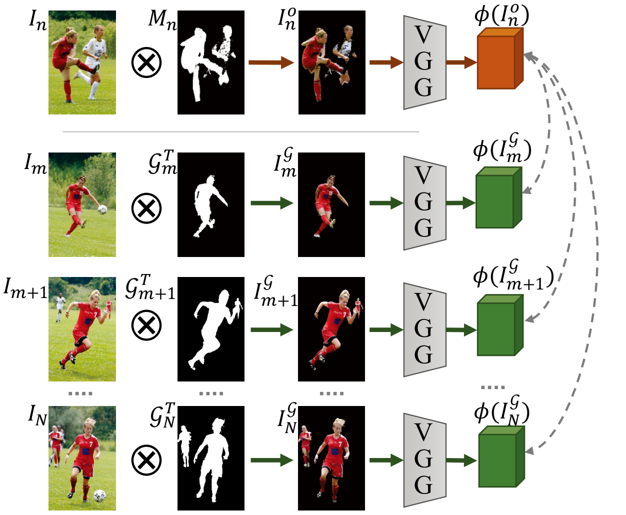

Figure 3: Illustration of co-perceptual loss L_c calculation.

4 Experiments

4.1 Experimental Setup

Datasets

We evaluated the proposed approach on three public benchmark datasets: iCoseg [Batra *et al.*, 2010], MSRC [Winn *et al.*, 2005] and Cosal2015 [Zhang *et al.*, 2016]. iCoseg consists of 38 groups of total 643 images which are challenging for co-saliency detection task because of the complex background and multiple co-salient objects, and the group size varies from 4 to 42. MSRC contains 7 groups of total 240 images, and each group has 30 $\sim$ 53 images. Cosal2015 is a more recent dataset and it has 50 groups and a total of 2015 images, each group containing 26 $\sim$ 52 images. Compared with iCoseg, MSRC and Cosal2015 have relative larger group size and are more challenging with various co-salient objects poses and sizes, greater appearance variations and even more complex backgrounds.

Implementation Details

We select the widely used pre-trained VGG 16-layer net [Simonyan and Zisserman, 2014] (over the MS COCO dataset [Lin *et al.*, 2014]) as the backbone network to extract the semantic features X_n for each image. The deconvolutional layers are initialized with simple bilinear interpolation parameters. For the sake of fair comparison, we following the same setting in [Wei *et al.*, 2017; Zheng *et al.*, 2018], the training groups are randomly selected from a subset of COCO dataset (which has 9213 images with pixel-wised ground-truth) using the global similarity (Gist and Lab color histogram features), and then fine-tune the model by randomly selecting $50\% - 50\%$ training-test images for three datasets. All images and groundtruth maps are resized to 224×224. The proposed models are optimized by standard SGD in which the momentum parameter is chosen as 0.99, the learning rate is set to 1e-5, and the weight decay is 0.0005. We need about 50000 training iterations for convergence. And for co-perceptual loss, we follow the setting in Neural Style Transfer and use the activated layers $Relu3\_1$,

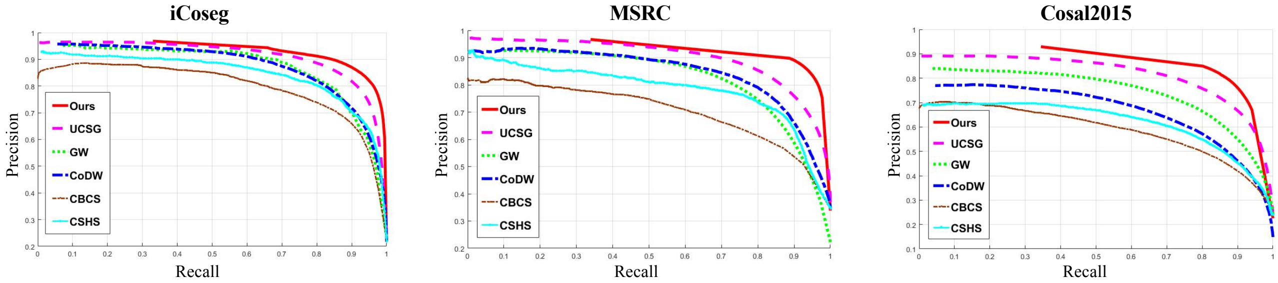

Figure 4: Comparison with the state-of-the-art methods with the same setting in terms of PR curves on three benchmark datasets.

Method	Setting	iCoseg			MSRC			Cosal2015		
		F_β	S_α	MAE	F_β	S_α	MAE	F_β	S_α	MAE
CBCS [Fu *et al.*, 2013]	CS+LF	0.690	0.671	0.166	0.587	0.496	0.300	0.514	0.545	0.234
CSHS [Liu *et al.*, 2014]	CS+LF	0.636	0.747	0.177	0.516	0.676	0.278	0.436	0.595	0.311
CoDW [Zhang *et al.*, 2016]	CS+DF	0.699	0.751	0.178	0.593	0.718	0.257	0.560	0.650	0.274
SP-MIL [Zhang *et al.*, 2017a]	CS+DF	0.703	0.782	0.159	0.625	0.775	0.212	-	-	-
GW [Wei *et al.*, 2017]	CS+DF	0.751	0.780	0.102	0.705	0.737	0.223	0.661	0.745	0.143
UCSG [Hsu *et al.*, 2018]	CS+DF	0.794	0.822	0.118	0.794	0.801	0.172	0.692	0.754	0.159
FASS [Zheng *et al.*, 2018]	CS+DF	0.838	0.867	0.062	0.808	0.804	0.144	0.696	0.802	0.109
Ours	CS+DF	0.877	0.908	0.033	0.870	0.846	0.076	0.751	0.835	0.076
DCL [Li and Yu, 2016]	SS+DF	0.799	0.851	0.073	0.801	0.783	0.151	0.692	0.763	0.135
Amulet [Zhang *et al.*, 2017b]	SS+DF	0.826	0.889	0.048	0.829	0.822	0.115	0.718	0.789	0.120
DSS [Hou *et al.*, 2019]	SS+DF	0.795	0.840	0.076	0.802	0.756	0.155	0.701	0.762	0.128

Table 1: Quantitative comparison with the state-of-the-arts on three famous benchmark datasets. SS and CS denote the single-image saliency and co-saliency methods, respectively. LF and DF indicate the conventional methods with low-level feature and deep learning based methods, respectively. The numbers in red and green respectively indicate the best and the second best results of the co-saliency methods.

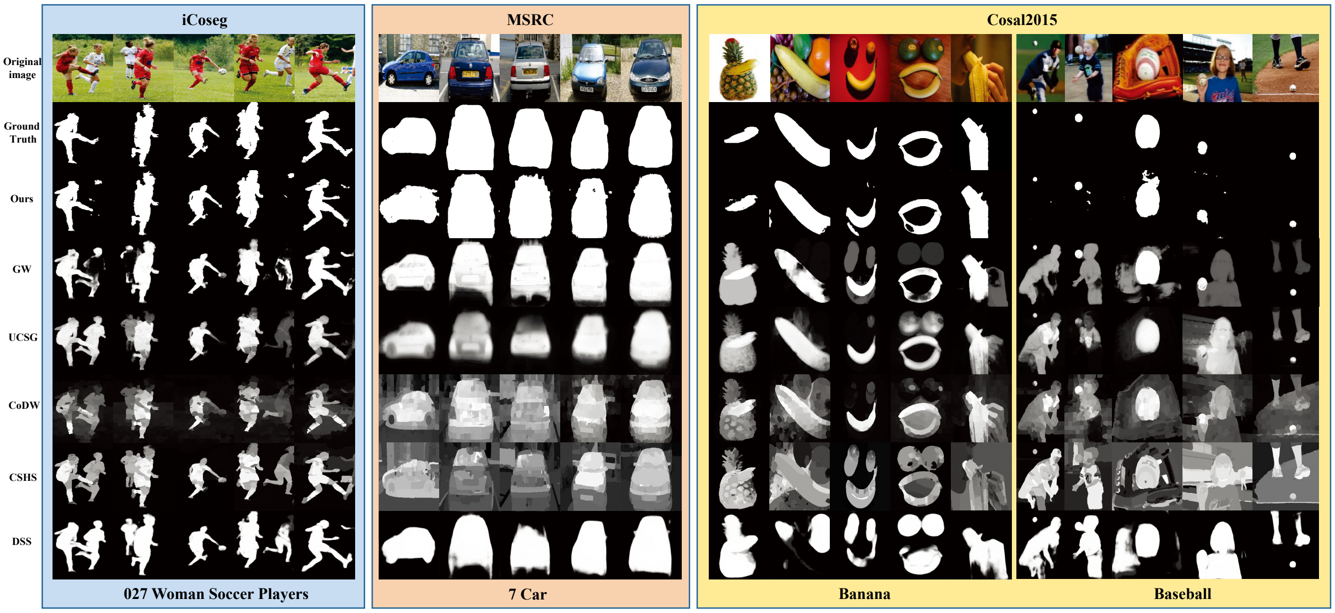

Figure 5: Visual comparison between the proposed method and the other representative methods on three benchmark datasets.

$Relu4\_1$, $Relu5\_1$ of VGG as the hidden representation. And the loss tradeoff parameter λ is set to be 0.1 in our work.

Evaluation Metrics

To evaluate the performance of the proposed method, four widely-used metrics are adopted: (1) Precision-Recall (PR) curve, which shows the tradeoff between precision and recall for different threshold (ranging from 0 to 255). (2)

F-measure (F_β) denotes the harmonic mean of the precision and recall values obtained by a self-adaptive threshold $T = \mu + \sigma$ (μ and σ are the mean value and standard deviation of co-saliency map.) The F_β is computed by $F_\beta = \frac{(1+\beta^2)\times Precision \times Recall}{\beta^2 \times Precision + Recall}$, where β^2 is typically set to 0.3 as suggested in [Han *et al.*, 2018; Yang *et al.*, 2013]. (3) Mean Absolute Error (MAE), which characterize the average 1-norm distance between ground truth maps and predictions. (4) Structure Measure (S_α) is adopted to evaluate the spatial structure similarities of saliency maps based on both region-aware structural similarity S_r and object-aware structural similarity S_o, defined as $S_\alpha = \alpha * S_r + (1 - \alpha) * S_o$, where $\alpha = 0.5$ [Fan *et al.*, 2017].

4.2 Comparison to the State-of-the-Arts

In order to evaluate the effectiveness of the proposed method, we compare it against 10 state-of-the-art algorithms. CSHS and CBCS are two conventional co-saliency approaches based on hand-crafted features that are widely compared in literatures. CoDW, SPMI, FASS and UCSG are 4 co-saliency methods using the deep learning features. GW is an end-to-end deep learning method for co-saliency detection. We also compare our method with 3 end-to-end single saliency methods which are DCL, Amulet and DSS. The experimental results are shown in Table 1. For fair comparison, we use either the implementations with recommended parameter settings or the saliency maps provided by the authors. As shown in Table 1, our approach outperforms all the state-of-the-art methods significantly in terms of all the metrics on three datasets. For example, for dataset Cosal2015, our method improves upon the second best algorithm FASS by about 8%, 4.1% and 30.3% in terms of F-measure, Structure Measure and MAE respectively. Note that although GW also had leveraged group feature in their work, our approach still exceeds GW by 13.6%, 12.1% and 46.8% in these three metrics. The same conclusion can be obtained from PR curves in Figure 4, the proposed method outperforms the state-of-the-arts by a large margin.

Figure 5 shows some sample co-saliency maps produced by the proposed approach and the state-of-the-art methods. From the results on Cosal2015 dataset, it can be observed that traditional method CSHS can hardly find common salient areas in complex cases of background clutter and cross-images variations. While the deep learning based single-image saliency method DSS and co-saliency methods UCSG and CoDW can better find the salient regions. But because of lacking group representation learning process, they all can't well handle the similarity between co-objects and non-common objects, resulting in undesired regions in co-saliency maps, like the white soccer players in iCoseg. The end-to-end deep learning method GW performs significantly better than others with less noise regions. However, our method produces the best saliency maps both in terms of the accuracy of contours and discrimination of different objects since we fully explore the single image properties and the robust group representation by RCAU and co-perceptual loss. Moreover, our results are more assertive than those of the others, which helps to easily select a binarization threshold to segment out the foregrounds given a co-saliency map.

4.3 Ablation Studies

In this section, we conduct evaluation on MSRC dataset to investigate the effectiveness of various components of the proposed model. The MSRC dataset has relative larger operable group size and is more challenging with various co-salient objects poses and sizes, greater appearance variations and complex backgrounds. The results are shown in Table 2. We set the baseline approach by only using single feature learning branch and training it with the cross-entropy loss L_s alone.

As can be seen, the baseline approach can't well handle the co-saliency task. After applying the RCAU, with the group representation, the performance of baseline approach is improved by 7.6%, 7.7% and 36.5% in terms of F-measure, Structure Measure and MAE respectively. This shows that the group representation is essentially crucial for co-saliency detection. When we replace our RCAU with the conventional recurrent units GRU, the performance drops a lot. That means our RCAU is more effective to capture the synergetic information in group compared with conventional recurrent units. However, the GRU still improves the performance of baseline, which means the recurrent architecture is naturally suitable for co-saliency detection task. We then add the co-perceptual loss L_c to form the completed version of our approach. The co-perceptual loss helps to further improve the performance comparing with RCAU version by 2.5%, 1.1% and 5% in terms of F-measure, Structure Measure and MAE. This indicates that the co-perceptual supervision can better model the synergetic relationships between the co-salient objects and help the network to learn more effective group representation, which in turn boost the co-saliency detection.

We evaluated the effects of RCAU with different component settings. As shown in Table 2, for update gate g_z, the combination of Z_s and Z_c achieves better performances than using them alone. When we remove the reset gate g_d from RCAU, the performance declines on three metrics especially on MAE. This indicates the reset gate g_d is able to suppress the noise information in the group. In order to verify that our recurrent architecture can handle different size image groups, we construct new testing groups by randomly selecting a sub-group with different size 5, 10 and 15 from the original groups. As reported in Table 2, our approach achieves good performance on different size groups and still consistently outperforms all the state-of-the-art methods. And the performance raises along with the group size, which emphasizes the importance of the group information completeness to robust co-saliency detection. To further justify the denoising ability of our RCAU, we add a noise image which is randomly selected from COCO dataset to the testing groups in size 5 and 10. As shown in results, although the noise data damages our performance a little, we still outperforms all the state-of-the-art methods, which demonstrate the robustness of the proposed method. As for other methods like GW, their performance declines badly under the influence of noise data.

5 Conclusion

In this paper, we propose a novel end-to-end deep learning approach with the recurrent co-attention network for robust

Method	MSRC		
	F_β	S_α	MAE
Baseline	0.789	0.781	0.126
Baseline+RCAU	0.849	0.841	0.080
Baseline+GRU	0.819	0.808	0.106
Baseline+RCAU+L_c	0.870	0.850	0.076
Baseline+L_c+RCAU(s)	0.862	0.849	0.076
Baseline+L_c+RCAU(c)	0.860	0.847	0.077
Baseline+L_c+RCAU(-g_d)	0.857	0.845	0.088
Baseline+RCAU+L_c (5)	0.851	0.841	0.081
Baseline+RCAU+L_c (10)	0.857	0.842	0.079
Baseline+RCAU+L_c (15)	0.862	0.846	0.076
Baseline+RCAU+L_c (5+noise)	0.842	0.833	0.085
Baseline+RCAU+L_c (10+noise)	0.853	0.836	0.080

Table 2: Ablation study of the proposed method on MSRC.

co-saliency detection. The proposed approach explores single image properties and robust group representation simultaneously, which are essentially crucial for co-saliency detection. By using the spatial and channel co-attention between images, the special designed RCAU recurrently explores all images in the group to learn the robust group representation and meanwhile suppresses noisy information. The group feature is then broadcasted to each individual image for boosting the inferring of co-saliency. Moreover, we propose a novel co-perceptual loss to make full use of interactive relationships of co-salient objects in the training group as the additional supervision. The whole modules are collaboratively optimized in an end-to-end manner, further improving the robustness of the approach. Extensive experimental results demonstrate the superiority of our approach.

Acknowledgments

This work was supported by National High Technology Re-search and Development Program of China (No. 2007AA01Z334), National Natural Science Foundation of China (Nos. 61321491 and 61272219), National Key Research and Development Program of China (Nos.2018YFC0309100, 2018YFC0309104), Program for New Century Excellent Talents in University of China (NCET-04-04605), The China Postdoctoral Science Foundation (Grant No. 2017M621700), National key research and development program of China (Nos.2018YFC0309100, 2018YFC0309104), and Innovation Fund of State Key Laboratory for Novel Software Technology (Nos. ZZKT2013A12ZZKT2016A11 and ZZKT2018A09).

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 *CVPR*, pages 3169–3176, 2010.

[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 *cvpr*, pages 2129–2136, 2011.

[Fan *et al.*, 2017] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4558–4567, 2017.

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

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

[Gatys *et al.*, 2016] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages pages 2414–2423, 2016.

[Han *et al.*, 2018] Junwei Han, Gong Cheng, Zhenpeng Li, and Dingwen Zhang. A unified metric learning-based framework for co-saliency detection. *IEEE Trans. Circuits Syst. Video Techn.*, 28(10):2473–2483, 2018.

[Hou *et al.*, 2019] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. *IEEE TPAMI*, 2019.

[Hsu *et al.*, 2018] Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, Xiaoning Qian, and Yung-Yu Chuang. Unsupervised cnn-based co-saliency detection with graphical optimization. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, pages 502–518, 2018.

[Li and Yu, 2016] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 478–487, 2016.

[Li *et al.*, 2017] Bo Li, Zhengxing Sun, Jiagao Hu, and Junfeng Xu. Co-saliency detection via sparse reconstruction and co-salient object discovery. In *Advances in Multimedia Information Processing - PCM 2017 - 18th Pacific-Rim Conference on Multimedia, Harbin, China, September 28-29, 2017, Revised Selected Papers, Part II*, pages 222–232, 2017.

[Li *et al.*, 2019] Bo Li, Zhengxing Sun, and Yuqi Guo. Supervae: Superpixelwise variational autoencoder for salient object detection. In *Proceedings of the Thirty-Three AAAI Conference on Artificial Intelligence, (AAAI-19), the 31th innovative Applications of Artificial Intelligence (IAAI-19), and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-19), Hawaii, Honolulu, USA, February*, 2019.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland,* *September 6-12, 2014, Proceedings, Part V*, pages 740–755, 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 Processing Letters*, 21(1):88–92, 2014.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[Wang *et al.*, 2018] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. Detect globally, refine locally: A novel approach to saliency detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3127–3135, 2018.

[Wei *et al.*, 2017] Lina Wei, Shanshan Zhao, Omar El Farouk Bourahla, Xi Li, and Fei Wu. Group-wise deep co-saliency detection. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3041–3047, 2017.

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

[Xue *et al.*, 2013] Jianru Xue, Le Wang, Nanning Zheng, and Gang Hua. Automatic salient object extraction with contextual cue and its applications to recognition and alpha matting. *Pattern Recognition*, 46(11):2874–2889, nov 2013.

[Yang *et al.*, 2013] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 3166–3173, 2013.

[Zhang *et al.*, 2016] 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.

[Zhang *et al.*, 2017a] Dingwen Zhang, Deyu Meng, and Junwei Han. Co-saliency detection via a self-paced multiple-instance learning framework. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(5):865–878, 2017.

[Zhang *et al.*, 2017b] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 202–211, 2017.

[Zheng *et al.*, 2018] Xiaoju Zheng, Zheng-Jun Zha, and Liansheng Zhuang. A feature-adaptive semi-supervised framework for co-saliency detection. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 959–966, 2018.