Robust Deep Co-Saliency Detection with Group Semantic

Chong Wang,[1,2] **Zheng-Jun Zha,**[2*] **Dong Liu,**[2] **Hongtao Xie**[2]
[1]Institute of Intelligent Machines, Chinese Academy of Sciences
[2]University of Science and Technology of China
wchong@mail.ustc.edu.cn, {zhazj, dongeliu, htxie}@ustc.edu.cn

Abstract

High-level semantic knowledge in addition to low-level visual cues is essentially crucial for co-saliency detection. This paper proposes a novel end-to-end deep learning approach for robust co-saliency detection by simultaneously learning high-level group-wise semantic representation as well as deep visual features of a given image group. The inter-image interaction at semantic-level as well as the complementarity between group semantics and visual features are exploited to boost the inferring of co-salient regions. Specifically, the proposed approach consists of a co-category learning branch and a co-saliency detection branch. While the former is proposed to learn group-wise semantic vector using co-category association of an image group as supervision, the latter is to infer precise co-salient maps based on the ensemble of group semantic knowledge and deep visual cues. The group semantic vector is broadcasted to each spatial location of multi-scale visual feature maps and is used as a top-down semantic guidance for boosting the bottom-up inferring of co-saliency. The co-category learning and co-saliency detection branches are jointly optimized in a multi-task learning manner, further improving the robustness of the approach. Moreover, we construct a new large-scale co-saliency dataset COCO-SEG to facilitate research of co-saliency detection. Extensive experimental results on COCO-SEG and a widely used benchmark Cosal2015 have demonstrated the superiority of the proposed approach as compared to the state-of-the-art methods.

1 Introduction

Discovering common and salient objects from a group of relevant images, termed as Co-saliency Detection, is beneficial for big image data management and various vision tasks, such as object co-segmentation (Quan et al. 2016; Dong et al. 2015; Zhang et al. 2016b), object co-localization(Tang et al. 2014), visual tracking (Li et al. 2018) and image retrieval (Fu, Cao, and Tu 2013; Zhang et al. 2013; Hong et al. 2017) etc.visual tracking (Li et al. 2018) and image retrieval (Fu, Cao, and Tu 2013; Zhang et al. 2013; Hong et al. 2017) etc. Co-saliency detection has attracted increasing interests from both academia and industry in recent years (Jeong, Hwang, and Cho 2017; Zhang, Meng, and Han 2017; Yao et al. 2017).

Effective visual representation characterizing salient and common objects is crucial for co-saliency detection. Conventional approaches utilize handcrafted features, such as color, texture and SIFT descriptors etc., and achieve encouraging performance (Fu, Cao, and Tu 2013; Liu et al. 2014; Li et al. 2015; Li, Meng, and Ngan 2013). However, handcrafted features based approaches usually suffer from multiple challenges including appearance variance of co-object across images, similar appearance between co-object and non-common object, and background clutter etc.Encouraged by the success of deep learning in many vision tasks (Krizhevsky, Sutskever, and Hinton 2012; Long, Shelhamer, and Darrell 2015; Zhang, Yu, and He 2018; Liu et al. 2016; Jiao et al. 2018; Xu et al. 2018), recent researches improve co-saliency detection by using deep neural network to learn visual representation in a data driven manner (Han et al. 2017; Zhang et al. 2016b; Zhang, Meng, and Han 2017; Zhang et al. 2016a; Jeong, Hwang, and Cho 2017). The deep visual features are fed into a subsequent co-saliency detection module. As the feature learning and co-saliency detection are separated as two independent processes, the learned features are not tailored for inferring co-salient regions, resulting in suboptimal performance. Recently, (Wei et al. 2017) proposed an end-to-end deep learning method for co-saliency detection, which integrates the process of feature learning and saliency mask prediction. A group-wise visual representation is designed to capture the interaction among visual features of individual images. As the group-wise visual feature is based on the concatenation of individual image features, it varies with the order of images within a group, limiting the robustness of the model.

Despite the remarkable progress made by recent works, they mainly focus on sophisticated inference of co-salient regions from visual cues, however neglect to explore high-level semantic supervision as well as inter-image interaction at semantic level, which are crucial for co-saliency detection. In this work, we propose a novel deep learning based approach for robust co-saliency detection. The proposed approach learns high-level group-wise semantic representation using inter-image common category association as supervision. The group-wise semantic representation characterizes the interaction among images at semantic level and is used as high-level semantic guidance for co-saliency inference. The group-wise semantic feature and co-salience map are jointly

[*]Corresponding author.

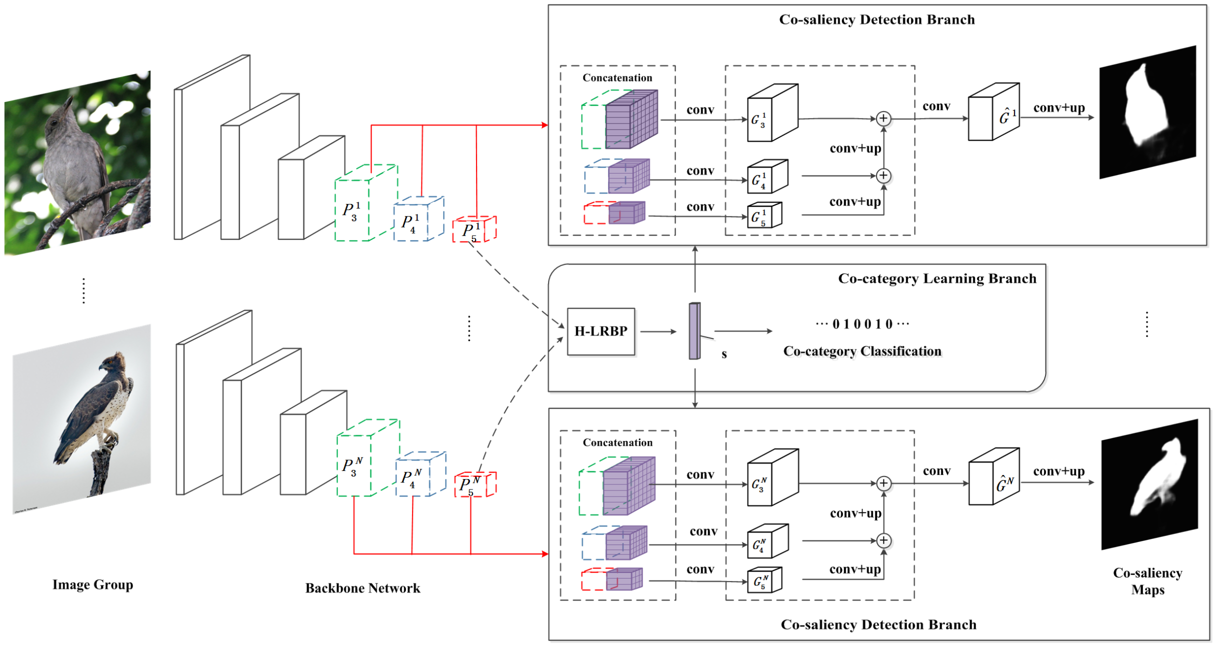


Figure 1: Overview of proposed robust co-saliency detection approach. A group of images $\{I_1, I_2, ..., I_N\}$ first go through the backbone networks to obtain the multi-scale feature maps $\{P_3^n, P_4^n, P_5^n\}, n \in [1, 2, ..., N]$. The feature maps $\{P_5^n\}_{n=1}^N$ from every images are aggregated to learn the group-wise semantic representation with the supervision of inter-image co-category association by the Co-category Learning branch. The group-wise semantic vector $\mathbf{s}$ is broadcasted to multi-scale visual features and is exploited as a top-down semantic guidance for boosting the inferring of co-salient regions by the Co-Saliency Detection branch.

optimized in an end-to-end learning manner. In particular, the proposed approach consists of two types of branches including *co-category learning branch* and *co-saliency detection branch* as shown in Figure 1. The *co-category learning branch* is proposed to learn group-wise semantic representation by using a Hierarchical Low-Rank Bilinear Pooling (H-LRBP) function with the supervision of co-category association, which endows the group-wise feature rich semantic clues related to the common categories of images. The *co-saliency detection branch* is designed to infer precise co-saliency maps by the joint exploration of both multi-scale deep visual features and high-level group-wise semantics. The group-wise semantic vector is broadcasted to each spatial location of multi-scale visual feature maps and is used as a top-down semantic guidance for boosting the bottom-up inferring of co-saliency. The multi-scale visual features augmented by group semantics are then assembled, providing a richer information flow path for the network. The complementarity and interaction of group semantics and multi-scale visual features are sufficiently exploited to facilitate the robust co-saliency reasoning. On the other hand, supervised co-saliency detection with deep learning is limited by the absence of large-scale co-saliency dataset. The largest existing dataset for co-saliency detection, Cosal2015 (Zhang et al. 2016b), consists of only 2,015 images with 50 groups. In order to facilitate the research of co-saliency detection, we construct a new large-scale dataset, i.e., COCO-SEG, selecting from the MS COCO2017 dataset (Lin et al. 2014). More details about the proposed dataset will be discussed in Section 3.1. Extensive experiments have been conducted on the COCO-SEG dataset and a widely used co-saliency dataset Cosal2015 . Experimental results have demonstrated that the proposed method outperforms the state-of-the-art approaches.

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:

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

where $F(\cdot)$ 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 a end-to-end learning scheme in this work. Inspired by the mechanism of human visual co-saliency, high-level semantic guidance as well as inter-image interaction at semantic level are important for co-saliency detection. Hence, we propose to learn a group-wise semantic representation $\mathbf{s} \in \mathbb{R}^D$ from the group of images:

$$\mathbf{s} = f_g(\mathcal{I}; \mathbf{\Theta}_g), \tag{2}$$

where $\mathbf{\Theta}_g$ represents model parameters. $\mathbf{s}$ is optimized using the common category association of images as supervision.

It is used as a high-level semantic guidance for inferring co-salient regions in each image:

$$M_n = f_{saliency}(I_n, \mathbf{s}; \mathbf{\Theta}_m), \tag{3}$$

where $\mathbf{\Theta}_m$ is the parameters of co-saliency prediction function. This work proposes a novel deep neural network architecture consisting of co-category learning branch and a co-saliency detection branch to jointly optimize the group-wise semantic feature learning and co-saliency map prediction in an end-to-end manner, as shown in Figure 1.

2.2 Co-category Learning

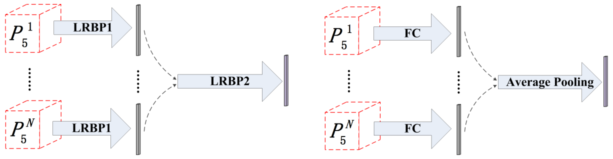
Co-category learning branch is proposed to learn group-wise semantic representation from a group of relevant images using inter-image co-category association as supervision. Inspired by the impressive achievements of bilinear pooling in fine-grained image classification task, it is believed that bilinear pooling can maintain selectivity in holistic features. Low-Rank Bilinear Pooling (LRBP) has been proposed in (Kong and Fowlkes 2017) to deal with the dimension explosion problem caused by full bilinear pooling. It is robust to feature redundancy and noise as well as saves the number of parameters and the cost of computing. Here, we design a hierarchical stack of pooling functions, termed as Hierarchical Low-Rank Bilinear Pooling (H-LRBP). H-LRBP first aggregates local features within a image feature map to a image representation and further encodes all image representations within a group into a holistic semantic representation. Such hierarchical pooling strategy decomposes the complicated mapping from individual image feature maps to group semantics and thus reduces the difficulty on learning group semantics.

H-LRBP: Denoting $X_n \in \mathbb{R}^{K \times D}$ as an image feature map which contains K local feature vectors, we defined the normalized bilinear pooling as $\frac{1}{K} X_n^\top X_n = \frac{1}{K}\sum_{i=1}^{K} {X_n^i}^\top X_n^i$. The C-way SVMs inference equation using the matrix representation of the bilinear feature is given as:

$$f^c(x_n) = \frac{1}{K} tr({W^c}^\top X_n^\top X_n) + b^c, \tag{4}$$

in which, $W^c \in \mathbb{R}^{D \times D}$ and b^c is the parameters of c-th SVM and $f^c(\cdot)$ is the corresponding predict function. Following (Kong and Fowlkes 2017), we impose a hard low-rank constraint on W_c by the parameterization with $U_+^c \in \mathbb{R}^{D \times r/2}$ and $U_-^c \in \mathbb{R}^{D \times r/2}$, namely $rank(W^c) = r$. Then, the SVM inference equation can be rewritten as:

$$\begin{aligned} f^c(X_n) &= \frac{1}{K}[tr(U_+^c {U_+^c}^\top X_n^\top X_n) - tr(U_-^c {U_-^c}^\top X_n^\top X_n)] + b_c \\ &= \frac{1}{K}[\|{U_+^c}^\top X_n^\top\|_F^2 - \|{U_-^c}^\top X_n^\top\|_F^2] + b_c. \end{aligned} \tag{5}$$



(a) H-LRBP (b) FC + Average Pooling

Figure 2: Two pooling strategies: (a) is the proposed H-LRBP and (b) is a baseline pooling compared in experiments.

Instead of the classification results, the pooled features are necessary for the subsequent processes in this work. Therefore, the prediction results of the C-way classifiers are used as C-dimensional features. A Hierarchical Low-Rank Bilinear Pooling (H-LRBP) strategy is proposed to aggregate the local visual features into group-wise representation $\mathbf{s}$ by using two consecutive LRBP layers in Eq.(5) as follows:

$$\mathbf{s} = f(\{f(X_n)\}_{n=1}^{N}), \tag{6}$$

where $f(\cdot)$ is the LRBP function. The pooling process is shown in Figure 2a. The group-wise semantic vector $\mathbf{s}$ is learned with the supervision of co-category association as elaborated in Section 2.4. The co-category learning only needs the co-category association among the images but does not require the awareness of what each category is. The co-category supervision endows $\mathbf{s}$ rich semantic clues related to the common categories of images and is valuable for facilitating the inferring of co-salient regions. Moreover, H-LRBP is not sensitive to the number and order of images within a group, making itself robust and flexible.

2.3 Co-saliency Detection with Group Semantic

Group-wise semantics represent high-level semantic cues and inter-image semantic interaction. We exploit the group-wise semantics as a top-down semantic guidance for boosting the bottom-up inferring of co-saliency. In particular, the group semantic vector is broadcasted to each spatial location of visual feature maps to augment visual features. The complementarity and interaction between high-level group semantics and low-level visual features are exploited to improve co-saliency detection. Regarding visual features, coarse-resolution features from high layers of neural network emphasize the abstraction of visual content and contains the context with large receptive field to the summary of object, while fine-resolution features from low layers emphasize the appearance details and are more conducive to the location of objects. In order to make full use of the complementary between multi-scale features, we fuse the visual features from multiple layers to provide a comprehensive representation for co-saliency mask prediction.

Taking VGG19 (Simonyan and Zisserman 2014) as the backbone ConvNet, we exploit the outputs of pool3, pool4 and pool5 layers as multi-scale feature maps, denoted as $\{P_3^n, P_4^n, P_5^n\}_{n=1}^{N}$ respectively. We concatenate the group-wise semantic vectors with each local feature vector, and a 1×1 convolutional layer is used to reduce the dimensionality. The resultant feature maps with group-wise semantic, called $\{G_3^n, G_4^n, G_5^n\}_{n=1}^{N}$, are jointly used for co-saliency detection as shown in Figure 1. Starting from G_5^n, the coarser-resolution feature map is upsampled by a factor of 2 using a deconvolutional layer. The upsampled map is then merged with the finer-resolution one by element-wise addition. This process iterates until the finest map $\hat{G}^n$ is obtained. To alleviate the aliasing effect of upsampling, we add a 3×3 convolutional layer after each merging operation. The final co-saliency maps $\mathcal{M}$ can be obtained with the fused features $\{\hat{G}^n\}$ by applying a 3×3 convolutional layer and a deconvolutional layer followed with a sigmoid activation function.

2.4 Loss Function

A classification loss is designed to learn group-wise semantic representation by the supervision of co-category supervision as follows:

$$\hat{y} = sigmoid(W^\top \mathbf{s} + b), \tag{7}$$

$$\mathcal{L}_{cls} = -\frac{1}{L}\sum_{l=1}^{L} y_l log\hat{y}_l - (1 - y_l)log(1 - \hat{y}_l), \tag{8}$$

where $W \in \mathbb{R}^{D \times L}$ and $b \in \mathbb{R}^L$ are parameters of L-class classifier. y_l and $\hat{y}_l$ are ground-truth and prediction value for the l-th co-category. $\mathcal{L}_{cls}$ is the classification loss function. We use sigmoid function as the activation function instead of softmax because there might exist more than one co-category appearing in an image group.

Denoting the ground-truth co-saliency masks of training image group as$\{\hat{M}_n\}_{n=1}^N$, the loss function of co-saliency detection is formulated as the average of pixel-wise cross entropy losses:

$$\mathcal{L}_{sal} = -\frac{1}{NP}\sum_{i=1}^{NP} \hat{M}_i log M_i - (1 - \hat{M}_i)log(1 - M_i) \tag{9}$$

where P is the pixel number of each training image. $\hat{M}_i$ and M_i are the i-th pixel values in ground truth and predicted co-saliency maps, respectively. Note that the co-category learning branch and the co-saliency detection branch are trained jointly, the overall loss function is given as

$$\mathcal{L} = \mathcal{L}_{sal} + \lambda \cdot \mathcal{L}_{cls}, \tag{10}$$

where λ is the tradeoff parameter.

3 Experiments

3.1 Datasets

The research and application of supervised co-saliency detection are limited by the lack of large-scale training data. The largest dataset for co-saliency detection at present, *i.e.,* Cosal2015 (Zhang et al. 2016b), consists of only 2,015 images in 50 groups. In order to evaluate the proposed approach as well as facilitate future research, we construct a new dataset, termed as COCO-SEG, at a large scale.

COCO-SEG is selected from the COCO2017 dataset (Lin et al. 2014) by applying the following selection strategies:

- All the images containing object of a certain category are grouped together, leading to 80 image groups. A group may contain more than one co-category.
- For a certain group, each image should belong to at least one co-category of the group with the region area over 4,000 pixels. This is to filter out images that contain only inconspicuous foreground.
- The groups containing less than 100 images in training set are removed to ensure that each group has enough training samples, leading to 78 groups finally.

The resultant COCO-SEG dataset contains 200K images belonging to 78 groups for training and 8K images of 78 groups for testing.

3.2 Implementation Details

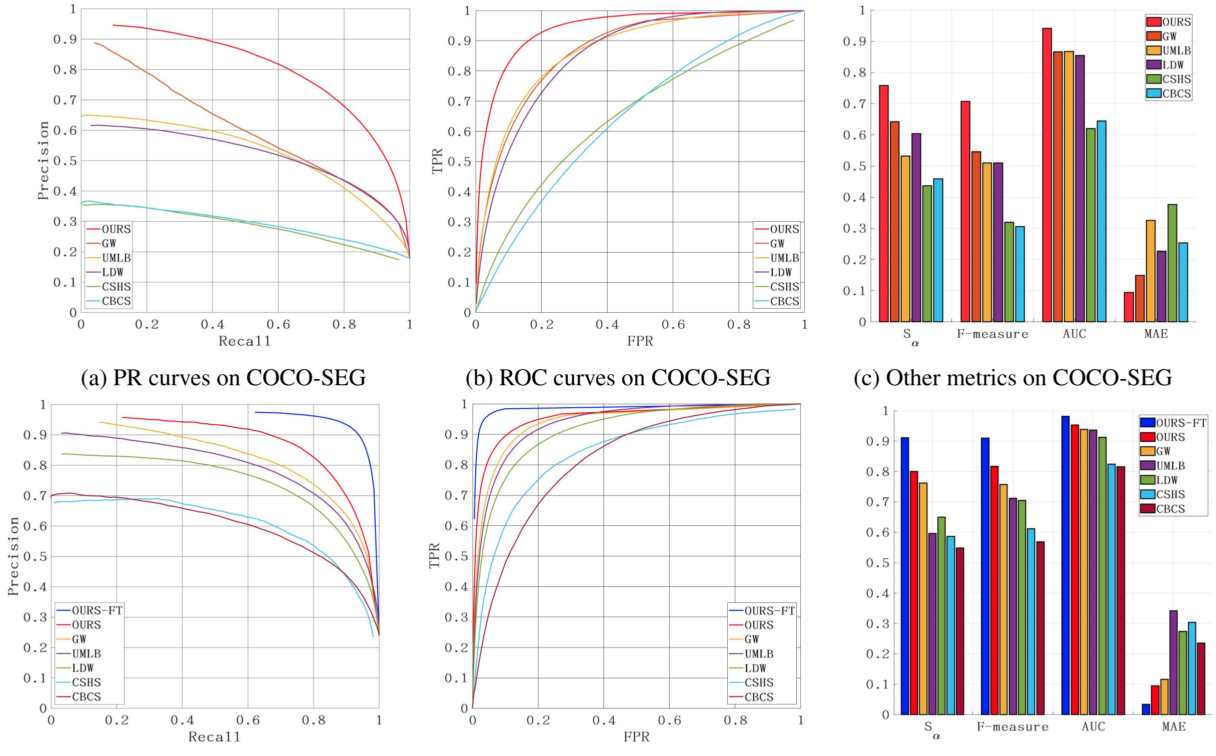
We select the widely used VGG 19-layer net (Simonyan and Zisserman 2014) as the backbone network for the sake of fair comparison, and initialize it with the parameters pre-trained for image classification task on ImageNet (Deng et al. 2009). The deconvolutional layers are initialized with simple bilinear interpolation parameters. We following the setting in (Wei et al. 2017), a sub-group consist of 5 images are randomly selected from a certain group, and a mini-batch consisting of 56 sub-groups are fed into the model at the same time during training. All images and ground-truth maps are resized to 224 × 224. The proposed models are optimized by the Adam Algorithm (Kingma and Ba 2014), in which the exponential decay rates for the first and second monent estimates are set to 0.9 and 0.999 respectively. The learning rate starts from 1e-5, and reduces by half every 10,000 steps until the model converges at about 50,000 steps. All the models are trained on the training set of COCO-SEG and evaluated on the COCO-SEG validation set and Cosal2015. During evaluation,we feed all images in a group into the model to generate all predicted masks simultaneously.

3.3 Evaluation Metrics

To evaluate the performance of the proposed method, six 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) Receiver Operating Characteristic (ROC) curve, which is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. (3) Area Under the Curve (AUC), which is the area under ROC curve. (4) Mean Absolute Error (MAE), which characterize the average 1-norm distance between ground truth maps and predictions. (5) F-measure (F_β), which is computed by:

$$F_\beta = \frac{(1 + \beta^2)Precision \times Recall}{\beta^2 Precision + Recall}, \tag{11}$$

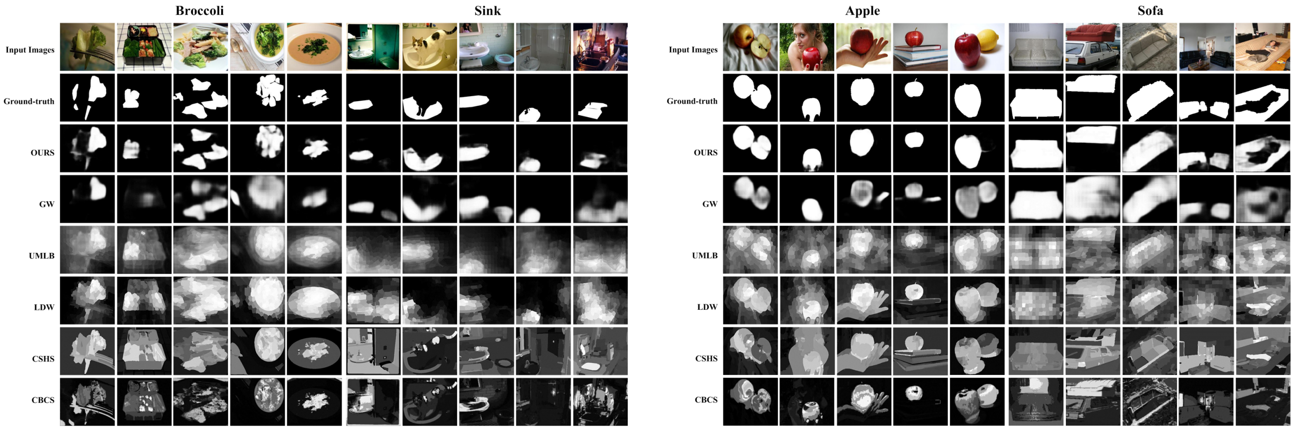
where the precision and recall are obtained by using a self-adaptive threshold $T = \mu + \varepsilon$ (μ and ε are the mean value and standard deviation of co-saliency map). β^2 is typically set to 0.3 as suggested in (Han et al. 2017; Yang et al. 2013). (6) Structure Measure (S_α) (Fan et al.



(a) PR curves on COCO-SEG (b) ROC curves on COCO-SEG (c) Other metrics on COCO-SEG

(d) PR curves on Cosal2015 (e) ROC curves on Cosal2015 (f) Other metrics on Cosal2015

Figure 3: Performance comparison between the proposed method and the state-of-the-art methods on COCO-SEG and Cosal2015 datasets



(a) COCO-SEG (b) Cosal2015

Figure 4: Visual comparison between the proposed method and the other representative methods on COCO-SEG dataset and Cosal2015 dataset

2017), which is a reliable metric to evaluate the spatial structure similarities of foreground maps. We set the hyper-parameter $\alpha = 0.5$ following (Fan et al. 2017).

3.4 Comparison to the State-of-the-Arts

In order to evaluate the effectiveness of the proposed method, we compare it against five state-of-the-art algorithms: GW(Wei et al. 2017), LDW(Zhang et al. 2016b), UMLB(Han et al. 2017), CSHS(Liu et al. 2014) and CBCS(Fu, Cao, and Tu 2013). GW is an end-to-end deep learning method for co-saliency detection. LDW and UMLB are two methods using the deep learning features. CSHS and CBCS are two conventional approaches based on hand-crafted features that are widely compared in literatures. For

Table 1: **Efficacy of co-category learning and H-LRBP**.

Metrics	Pooling	$\lambda = 0$	$\lambda = 0.1$	$\lambda = 1$
S_α	AP	0.7070	0.7331	0.7473
	H-LRBP	0.7413	0.7492	**0.7578**
F_β	AP	0.6430	0.6801	0.6994
	H-LRBP	0.6848	0.7029	**0.7051**
AUC	AP	0.9157	0.9252	0.9317
	H-LRBP	0.9377	0.9348	**0.9398**
MAE	AP	0.1217	0.1016	0.0945
	H-LRBP	0.1048	0.0939	**0.0927**

the sake of fair comparison, VGG-19 is used as backbone networks for GW, LDW and UMLB. Figure 3a - 3c illustrate the experimental results in terms of various metrics on COCO-SEG dataset. Our approach outperforms the state-of-the-art methods significantly in terms of all the metrics. For example, our method improves upon the second best algorithm GW by about 18%, 26%, 7% and 35% in terms of Structure Measure, F-measure, AUC and MAE respectively.

The proposed method is also evaluated on a widely used benchmark co-saliency dataset Cosal2015. As illustrated in Figure 3d - 3f, the proposed method achieves better performance than existing ones in terms of all the metrics. For example, the proposed method improves upon the second best method GW by about 5%, 6% and 13% in terms of Structure Measure, F-measure and MAE respectively. As aforementioned, all the models are trained on COCO-SEG and applied on Cosal2015. The experimental results have shown that the proposed method can obtain consistent performance improvements when dealing with new categories of foregrounds that do not appear during training. Thus, the proposed method possess better generalization ability and flexibility in practical use. We also fine-tune the model without the optimization of co-category learning branch in Cosal2015 by selecting 50%-50% training-test images as (Han et al. 2017). The fine-tuned model is denoted as "OURS-FT", which exhibits impressive performance as shown in Figure 3d - 3f.

Figure 4 shows some sample co-saliency maps produced by the proposed approach and the state-of-the-art methods on COCO-SEG and Cosal2015 datasets. From the results on COCO-SEG dataset, it can be observed that two traditional methods CSHS and CBCS can hardly find common foreground areas in complex cases of high inter-class similarity and intra-class variations. While the end-to-end deep learning method GW performs significantly better than the traditional methods, our method produces the best saliency maps both in terms of the accuracy of contours and discrimination of different objects. From the saliency maps on Cosal2015 dataset, it can be seen that our method is more robust to the inter-class similarity and intra-class variations and is more meticulous in shaping the appearance details. Moreover, co-saliency maps produced by our method 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.

Table 2: **Efficacy of multi-scale features**: With $\lambda = 0$, the models using multi-scale features are compared with the models using single-scale features.

Metrics	AP single-scale	AP multi-scale	H-LRBP single-scale	H-LRBP multi-scale
S_α	0.6854	0.7070	0.6951	**0.7413**
F_β	0.6127	0.6430	0.6248	**0.6848**
AUC	0.9010	0.9157	0.9104	**0.9377**
MAE	0.1245	0.1217	0.1208	**0.1048**

Table 3: **Comparison between group-wise representation**: With $\lambda = 0$ and single-scale features, the models using AP and H-LRBP are compared against the existing GW method.

Metrics	GW	AP	H-LRBP
S_α	0.6424	0.6854	**0.6951**
F_β	0.5455	0.6127	**0.6248**
AUC	0.8665	0.9010	**0.9104**
MAE	0.1494	0.1245	**0.1208**

3.5 Ablation Studies

In this section, we conduct evaluation to investigate the effectiveness of various components of the proposed model. The models using visual feature from multiple layers and single layer are denoted as "multi-scale" and "single-scale", respectively. For pooling strategy, "AP" refers to the "FC + Average Pooling" strategy as show in Figure 2b, which uses fully connected layer to integrate local features into image representations and then pools all image features within a group into group-wise representation by average-pooling. All the ablation studies are conducted on the proposed COCO-SEG dataset.

Efficacy of co-category learning and H-LRBP Table 1 provides the performance of the proposed method with various values of λ and the two pooling strategies. λ is the trade-off parameter adjusting the relative strength of exploiting co-category association supervision in learning group-wise representation. From the results, we can obtain the following observations: (1) the models with λ as 0.1 and 1 outperforms the mode with λ as 0. This indicates that the exploration of co-category supervision is able to learning more effective group-wise representation with semantic cues, which in turn boost the co-saliency detection. The model with λ as 1 obtains the best performance; (2) "H-LRBP" achieves consistent performance improvements over "AP" in various settings.

Efficacy of multi-scale visual features In order to better investigate the effectiveness of multi-scale visual features, we compare the models using single-scale and the multi-scale features in the setting of non-supervision of co-category association, *i.e.,* $\lambda = 0$. The single-scale feature is processed by applying a convolutional layer followed with a deconvolutional layer on $\{G_5^n\}_{n=1}^N$. As illustrated in Table 2, the performance of multi-scale visual features is much better than that of single-scale features with either "AP" or "H-LRBP" strategy. The results have demonstrated that fusing visual features at multi-scales produces a comprehensive representation characterizing both visual abstraction and details of foregrounds and is useful for co-saliency prediction.

Comparison between group-wise features We compare the proposed model without co-category supervision and multi-scale features to the existing method GW (Wei et al. 2017) which learns a group-wise visual representation for co-saliency detection. As shown in Table 3, both the model with "AP" and "H-LRBP" strategies outperform GW significantly. This indicates that the proposed method is able to learn more effective group-wise representation even without the supervision of co-category association. Moreover, as the group-wise feature in GW is based on the concatenation of individual image features, it varies with the order of images within a group. Our method uses H-LRBP for feature agreation. It is not affected by the number and order of images within a group and thus possess better robustness and flexibility.

4 Conclusion

This paper proposed a new deep learning based approach for robust co-saliency detection, consisting of a co-category learning branch and a co-saliency detection branch. The proposed approach explores the high-level semantic supervision as well as inter-image interaction at semantic level, which are important for co-saliency detection. A group-wise semantic representation characterizing inter-image semantic interaction is learned by the proposed co-category learning branch using the co-category association as supervision. The co-saliency detection branch infers precise co-saliency maps by using the group semantic as a top-down semantic guidance as well as visual features at multiple scales. The group-wise semantic feature and co-salience map are jointly optimized in an end-to-end multi-task learning manner. Moreover, we constructed a new large-scale co-saliency dataset COCO-SEG, which is the largest dataset for co-saliency detection at present. Extensive evaluation on both COCO-SEG and the benchmark Cosal2015 have demonstrated that the proposed approach outperforms multiple state-of-the-art methods in terms of various performance matrics.

5 Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 61622211, 61472392 and 61620106009 as well as the Fundamental Research Funds for the Central Universities under Grant WK2100100030.

References

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Dong, X.; Shen, J.; Shao, L.; and Yang, M.-H. 2015. Interactive cosegmentation using global and local energy optimization. *IEEE Transactions on Image Processing* 24(11):3966–3977.

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision*, 4558–4567.

Fu, H.; Cao, X.; and Tu, Z. 2013. Cluster-based co-saliency detection. *IEEE Transactions on Image Processing* 22(10):3766–3778.

Han, J.; Cheng, G.; Li, Z.; and Zhang, D. 2017. A unified metric learning-based framework for co-saliency detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Hong, R.; Li, L.; Cai, J.; Tao, D.; Wang, M.; and Tian, Q. 2017. Coherent semantic-visual indexing for large-scale image retrieval in the cloud. *IEEE Transactions on Image Processing* 26(9):4128–4138.

Jeong, D.-j.; Hwang, I.; and Cho, N. I. 2017. Co-salient object detection based on deep saliency networks and seed propagation over an integrated graph. *arXiv preprint arXiv:1706.09650*.

Jiao, Y.; Li, Z.; Huang, S.; Yang, X.; Liu, B.; and Zhang, T. 2018. 3d attention-based deep ranking model for video highlight detection. *IEEE Transactions on Multimedia*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kong, S., and Fowlkes, C. 2017. Low-rank bilinear pooling for fine-grained classification. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 7025–7034. IEEE.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Li, Y.; Fu, K.; Liu, Z.; and Yang, J. 2015. Efficient saliency-model-guided visual co-saliency detection. *IEEE Signal Processing Letters* 22(5):588–592.

Li, Z.; Zhang, J.; Zhang, K.; and Li, Z. 2018. Visual tracking with weighted adaptive local sparse appearance model via spatio-temporal context learning. *IEEE Transactions on Image Processing*.

Li, H.; Meng, F.; and Ngan, K. N. 2013. Co-salient object detection from multiple images. *IEEE Transactions on Multimedia* 15(8):1896–1909.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, Z.; Zou, W.; Li, L.; Shen, L.; and Le Meur, O. 2014. Co-saliency detection based on hierarchical segmentation. *IEEE Signal Processing Letters* 21(1):88–92.

Liu, J.; Zha, Z.-J.; Tian, Q.; Liu, D.; Yao, T.; Ling, Q.; and Mei, T. 2016. Multi-scale triplet cnn for person re-identification. In *Proceedings of the 2016 ACM on Multimedia Conference*, 192–196. ACM.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Quan, R.; Han, J.; Zhang, D.; and Nie, F. 2016. Object co-segmentation via graph optimized-flexible manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 687–695.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tang, K.; Joulin, A.; Li, L.-J.; and Fei-Fei, L. 2014. Co-localization in real-world images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1464–1471.

Wei, L.; Zhao, S.; Bourahla, O. E. F.; Li, X.; and Wu, F. 2017. Group-wise deep co-saliency detection. *arXiv preprint arXiv:1707.07381*.

Xu, Y.; Han, Y.; Hong, R.; and Tian, Q. 2018. Sequential video vlad: training the aggregation locally and temporally. *IEEE Transactions On Image Processing* 27(10):4933–4944.

Yang, C.; Zhang, L.; Lu, H.; Ruan, X.; and Yang, M.-H. 2013. Saliency detection via graph-based manifold ranking. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 3166–3173. IEEE.

Yao, X.; Han, J.; Zhang, D.; and Nie, F. 2017. Revisiting co-saliency detection: a novel approach based on two-stage multi-view spectral rotation co-clustering. *IEEE Trans. Image Process* 26(7):3196–3209.

Zhang, H.; Zha, Z.-J.; Yang, Y.; Yan, S.; Gao, Y.; and Chua, T.-S. 2013. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *Proceedings of the 21st ACM international conference on Multimedia*, 33–42. ACM.

Zhang, D.; Han, J.; Han, J.; and Shao, L. 2016a. Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining. *IEEE transactions on neural networks and learning systems* 27(6):1163–1176.

Zhang, D.; Han, J.; Li, C.; Wang, J.; and Li, X. 2016b. Detection of co-salient objects by looking deep and wide. *International Journal of Computer Vision* 120(2):215–232.

Zhang, D.; Meng, D.; and Han, J. 2017. Co-saliency detection via a self-paced multiple-instance learning framework. *IEEE transactions on pattern analysis and machine intelligence* 39(5):865–878.

Zhang, W.; Yu, X.; and He, X. 2018. Learning bidirectional temporal cues for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology* 28(10):2768–2776.