ELR-Net: An Enhancement and Location-Refinement Network for Co-Saliency Detection

Haiyang He
College of Computer and Information Sciences, Fujian Agriculture and Forestry University
Fuzhou, China
1464751753@qq.com

Xiaolin Li
College of Computer and Information Sciences, Fujian Agriculture and Forestry University
Fuzhou, China
li1180@sina.com.cn

Zhihong Wang
College of Computer and Information Sciences, Fujian Agriculture and Forestry University
Fuzhou, China
1084472397@qq.com

Shiguo Huang*
College of Computer and Information Sciences, Fujian Agriculture and Forestry University
Fuzhou, China
446588040@qq.com

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671666

***Abstract*—Co-salient object detection is a challenging task, which aims to segment the co-occurring salient objects in multiple images at the same time. To address this task, we propose an end-to-end Enhancement Location-Refinement Network (ELR-Net) to capture both salient and repetitive visual patterns from multiple images. For various scenarios, common objects in different images only have the same semantic information, so we first propose a deep co-salient method based on channel and spatial attention module (CASM), which combines the attention mechanism to enhance the common semantic information. Subsequently, we employ a Co-attention and Refinement Module (CARM) to capture the common attributes of co-salient objects by learning the features consensus representation from a group of images using our group affinity module (GAM) and then we develop a self-correlation module (SCM) to further fine-grained information on the co-salient regions. Specifically, SCM can maintain the feature independence upon semantic categories and further help our model to distinguish pixels with similar but different categories. Moreover, single image saliency maps (SISMs) are predicted to extract intra-saliency cues, and then a correlation fusion module (CFM) is employed to extract inter-saliency cues. The proposed ELR-Net is evaluated on three challenging benchmarks, i.e., CoSal2015, CoSOD3k, and CoCA, demonstrate that our ELR-Net outperforms 9 cutting-edge models and achieves state-of-the-art performance.**

***Index Terms*—co-salient object detection, attention mechanism, location and refinement**

I. Introduction

Salient object detection (SOD) is an important topic in the field of computer vision [1–3]. SOD aims to extract the most attractive regions from a single image. Co-salient object detection (CoSOD) is a special type of SOD and its goal is to segment common salient objects from two or more images at the same time. It can be applied to various computer vision tasks, such as co-segmentation [4, 5], semantic segmentation [6], video analysis/co-localization [7, 8], etc.

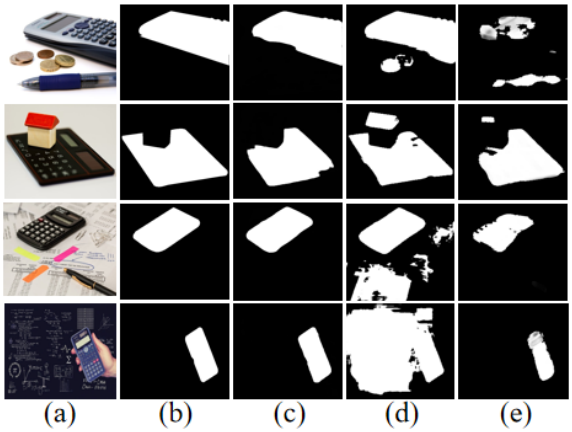

Fig. 1. Comparison with state-of-the-art methods in complex real-world scenarios. (a) Image; (b) Ground truth; (c) ELR-Net (ours); (d) GCoNet; (e) CoEGNet

Early CoSOD methods use handcrafted features to explore the consistency among multiple images, e.g., SIFT [8], color [9], texture [10, 11], or multiple cues fusion [12]. However, the discrimination of these shallow features is not enough to segment co-salient objects of real-world scenarios. Recently, deep-learning-based methods [13, 14] have greatly improved the performance of the CoSOD task by exploring the semantic correlations within a group of images. Later, [15, 16] proposed

* Corresponding author

the end-to-end deep learning frameworks to integrate the process of feature learning and saliency map prediction. The methods of [17–19] learned intra and inter saliency cues from similar foregrounds within an image group. The works of [20, 21] took the image saliency maps to exploit the intra saliency cues and achieve comparable results. [22] introduced a novel group collaborative learning strategy to learn intra and inter saliency cues. Although some co-salient methods have been proposed in the last few years and make remarkable progress, challenges still exist for further research. Previous methods do not investigate the deep guidance of high-level features. In addition, for mining the co-relationship, most existing CoSOD frameworks focus mainly on the region's accuracy but neglect the fine-grained information.

To address the aforementioned challenges, in this paper, we propose an ELR-Net for fine-grained CoSOD performance. Our ELR-Net consists of three components: Encoder Feature Correlation, Location-Refinement, Intra and Inter cues Correlation. Feature Correlation Encoder component contains a channel and spatial attention module (CASM). Location-Refinement component contains co-attention and refinement module (CARM). CARM contains two sub-modules i.e. group affinity module (GAM) and self-correlation module (SCM). Intra and Inter cues Correlation module first utilizes single image saliency maps (SISMs) and deep features to explore intra cues. And then to explore the inter cues, a correlation fusion module (CFM) is employed to extract inter-saliency through capturing the correlations between single image features and the extracted intra cues. As shown in Fig.1, the proposed method can handle some challenging scenarios. We can observe that (1) our model can group the co-salient objects (calculator) effectively through exploring correlations from a group of images (see the first row); (2) our model can eliminate the influence of occlusions (see the second row); (3) our model can filter the background clutters (see the last two rows).

From the mentioned above, the main contributions of this paper are four-fold:

- We design an Enhancement and Location-Refinement Network, which consists of Encoder Feature Correlation, Location-Refinement, Intra and inter cues Correlation to address the CoSOD problem. Ablation studies validate its effectiveness.
- We present a Co-attention and Refinement Module (CARM) to locate the co-occurring salient regions. GAM first enables the model to concentrate on the co-salient regions through exploring correlations from a group of images. Then, SCM is used to further refine the co-salient regions by maintaining the feature independence upon semantic categories.
- We introduce a channel and spatial attention module (CASM) to enhance the high-level semantic features from the backbone network to obtain the co-features, which have common objects and highlight the location information of objects. To our best knowledge, this is the first to introduce the attention mechanism to excavate common information cues from a group of high-level semantic features in the CoSOD task.
- Extensive experiments on three challenging CoSOD benchmarks, i.e., CoSal2015, CoSOD3k, and CoCA, show that our ELR-Net achieves state-of-the-art performance.

II. Proposed Enhancement and Location-Refinement Network

A. Overall Network Architecture

Given a group of N relevant images $I = \{I_i\}_{i=1}^n$, CoSOD aims to segment their common salient object(s) and generate the corresponding co-saliency maps $M = \{M_i\}_{i=1}^n$. Figure 2 illustrates the flowchart of our ELR-Net. First, the image group I is fed into an encoder network to exact l_2-normalized high-level semantic features F=$\{F_i\}_{i=1}^n$. Channel and spatial attention module (CASM) is used to enhance the F to obtain the common information cues, denoted as F_h. Co-attention and Refinement Module (CARM) is applied to concentrate on the co-salient regions and further fine-grained information to produce F_g. We integrate single image saliency maps (SISMs), denoted as $S = \{S_i\}_{i=1}^n$ combine with F_h to obtain intra cues F_c, we further employ a correlation fusion module (CFM) to exploit correlations between F_c and F_h, and generate inter cues F_e. Finally, F_e and F_g are fed into a decoder network to predict the co-saliency maps M.

B. Channel and spatial attention module

It is well known that different feature channels correspond to different semantic information. Therefore, we adopt a channel and spatial attention module to enhance the common information between a group of input images. The intuitive presentation of CASM can be seen in Figure 3. Given a group of relevant images, objects of the same classes have the activated value in the same feature channel [23]. Channel attention focuses on which common channels of the input images are activated, and enhances the features of these common channels. Spatial attention enhances the local spatial information and suppresses the background noise information through focusing on which space of the feature maps is activated. The operation of the CASM is defined as Equation (1).

$$F_h = S_{att}(C_{att}(F)) \tag{1}$$

where F denotes the high-level semantic features from the backbone network. $C_{att}(\cdot)$ denotes the channel attention, $S_{att}(\cdot)$ denotes the spatial attention. $C_{att}(\cdot)$ is formulated as Equation (2).

$$C_{att}(F) = G(A_{max}(F)) \otimes F \tag{2}$$

where A_{max} represents the adaptive max-pooling operation, $G(\cdot)$ is a two-layer perceptron, and $\otimes$ denotes the multiplication by the dimension broadcast. $S_{att}(\cdot)$ is formulated as Equation (3).

$$S_{att}(F) = Conv(M_{max}(F)) \otimes F \tag{3}$$

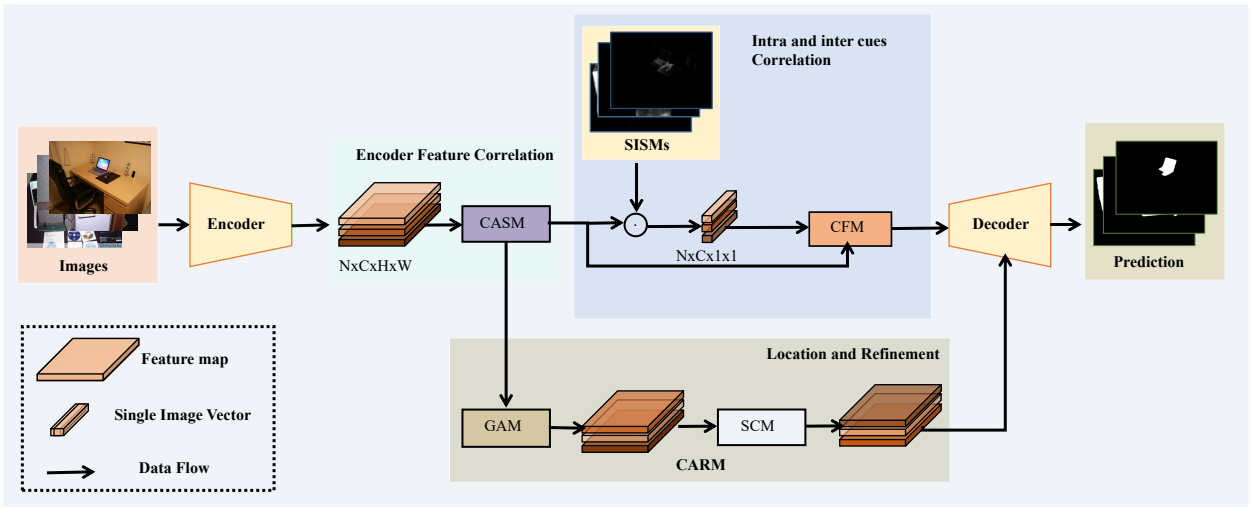

Fig. 2. **Pipeline of proposed ELR-Net.** We assume the input image group consists of three images to simplify illustration. We first utilize CASM to enhance the high-level semantic features from Encoder (VGG16). Then we employ the CARM to locate and refine co-saliency features, denoted as F_g. Inter and intra cues correlation is used to explore intra and inter cues by correlation techniques, denoted as F_c and F_e, respectively. Finally, F_e and F_g are integrated to generate co-saliency maps.

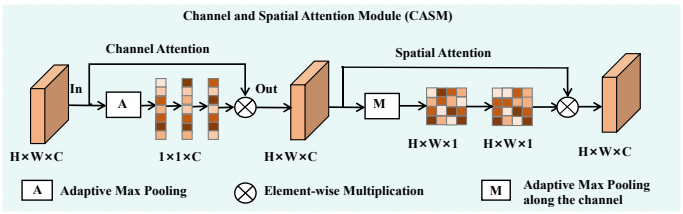

Fig. 3. Architecture of Channel and Spatial Module (CASM).

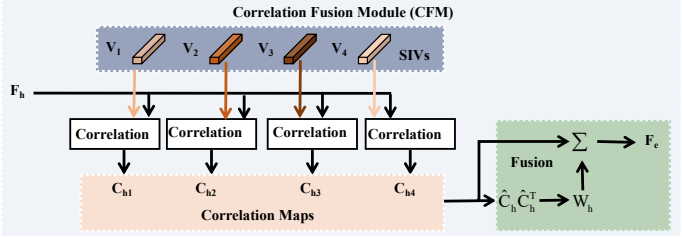

Fig. 4. Architecture of Correlation Fusion Module (CFM). "Σ" denotes weighted summation. For simplicity, here we only use 4 images for simplicity and better illustration.

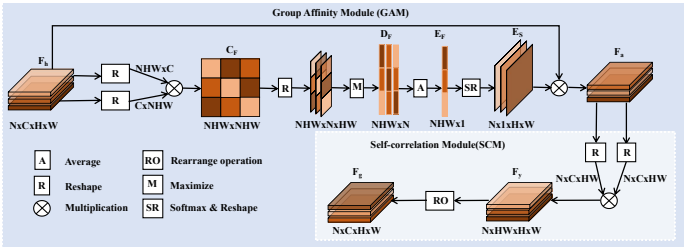

Fig. 5. Co-attention and Refinement Module (CARM).

where $M_{max}(\cdot)$ is the adaptive max-pooling operation along the channel. The CASM is different from the operation proposed in [24], which excavates informative cues from depth features. Our CASM aims to enhance the common features cues and the local spatial information from a group of images. To our best knowledge, we are the first to introduce a deep co-salient method based on channel and spatial attention mechanism to excavate common cues from a group of images in the CoSOD. Besides, we only leverage a single adaptive max-pooling operation to excavate the common cues of a group of images in the CASM and reduce the complexity of the module simultaneously, which is based on the intuition that RGBD aims at excavating the most critical cues in depth features.

C. Extraction of Intra and Inter Cues

Intra and inter cues are vital for locating the regions in the CoSOD task. Inspired by [25], we directly use SISMs and semantic features with normalized masked average pooling (NMAP) operation [26] into our network for end-to-end training to explore more discriminative intra cues. As shown in Figure 2, given a group of image features F_h=$\{F_{hi}\}_{i=1}^{n}$($F_{hi} \in \mathbb{R}^{C\times H\times W}$) after CASM, we adjust the corresponding SISMs $S = \{S_i\}_{i=1}^{n}$ to generate single-image vectors ($SIVs$) V=$\{V_i\}_{i=1}^{n}$ and they can be calculated by Equation (4).

$$\begin{cases} \hat{v_i} = \dfrac{1}{HW}\displaystyle\sum_{x=1}^{H}\sum_{y=1}^{W} F_i(:,x,y) \odot S_i(:,x,y) \\ v_i = \dfrac{\hat{v_i}}{||\hat{v_i}||_2} \end{cases} \tag{4}$$

where $\odot$ denotes the element-wise multiplication, x and y are the indices along the spatial dimensions, $||\cdot||_2$ is the l_2 norm. In our network, SISMs are directly used to filter out

the features of non-salient regions by multiplication, rather than as the training targets to force the CoSOD models into overfitting inaccurate SISMs with a performance drop. Then, NMAP is used to obtain $v_i \in \mathbb{R}^C$, which can express latent intra-saliency categories. In this way, even though the SISMs are not accurate, the inaccurate features in the maps can be largely diluted after NMAP.

To extract inter cues from the intra ones, we simply embed the Correlation Fusion Module (CFM) into our network. The details can be seen in Figure 4. We take the h-th image feature F_h from a group of N images as an example. Note that we set N=4 for simplicity and better illustration. For each $SIVs$ v_i, we compute the inter-product between it and pixel-wise feature vectors in F_h to obtain correlation map $C_{hi} \in \mathbb{R}^{H \times W}$. Each C_{hi} address the region of F_h that has a high response to the intra-saliency category due to $SIVs$ v_i. Then, we fuse $\{C_{hi}\}_{i=1}^n$ with a weight vector to obtain a matrix $\hat{C}_h \in \mathbb{R}^{n \times HW}$, which considers the relevance between each pair of correlation maps. The above process can be described as Equation (5).

$$W_h = softmax(\alpha \hat{C}_h \hat{C}_h^T 1) \tag{5}$$

where α is a learned parameter to adjust the vector to proper magnitude for the following softmax normalization, $\hat{C}_h \hat{C}_h^T \in \mathbb{R}^{n \times n}$ is a correlation matrix, which represents the relevance between every pair correlation maps, which 1 represents an n-dimensional vector of all ones. Finally, with the weight vector W_h, we sum the correlation maps $\{C_{hi}\}_{i=1}^n$ followed by a min-max normalization and extract the feature map F_e as the inter cues.

D. Co-attention and Refinement Module

In this subsection, we introduce a novel Co-attention and Refinement Module (CARM) to explore co-salient features among all images in a group, which is based on the intuition that common objects from the same class always have some similar appearance and high similarity in their corresponding features. Self-supervised video tracking methods [27–30] ultilize the pixel-wise correspondences between two adjacent frames in the videos to obtain segmentation masks of target objects. So, we extend this idea and combine it with some existing co-salient methods [22, 25] to help our model locate and further refine the co-salient regions. The details are shown in Figure 5.

Our CARM is divided into two parts including the group affinity module (GAM) and self-correlation module (SCM). GAM can capture the common attributes of co-salient objects in an image group (Location) and SCM further fine-grained information on the co-salient regions (Refinement). Specifically, given the feature maps $F_h \in \mathbb{R}^{N \times C \times H \times W}$, we shape it into the size of $NHW \times C$ and $C \times NHW$. Then we calculate the co-attention map $C_F \in \mathbb{R}^{NHW \times NHW}$. For the C_F ,we further find the maxima for each image $D_F \in \mathbb{R}^{NHW \times N}$ from C_F, then generate the global co-attention map $E_F \in \mathbb{R}^{NHW \times 1}$ through averaging all the maxima of N images, in this way, E_F can alleviate the influence of occasional co-occurring bias through globally optimized on all images. Then, we use a softmax operation to normalize E_F to obtain the $E_S \in \mathbb{R}^{N \times 1 \times H \times W}$ and we multiply E_S with the original feature F_h to produce the $F_a \in \mathbb{R}^{N \times C \times H \times W}$, which focuses on capturing the commonality among co-salient objects. However, we observe that in this way our network will make sub-optimal predictions that fail to distinguish pixels with similar but different categories. We further fine-grained information on the co-salient regions. Specifically, we shape F_a into the size of $N \times C \times HW$, denoted as $\hat{F}_a$. Then we calculate the self-correlation matrix $\hat{F}_a^T \hat{F}_a \in \mathbf{R}^{N \times HW \times HW}$, and reshape it into the size of $N \times HW \times H \times W$ to obtain the F_y, then we combine it with F_e (the result of intra and inter cues module) to rearrange the channel order of F_y to obtain F_g. In this way, the channel order of F_g is independent of the pixel positions. Specifically, for the pixel (x,y) of the higher co-saliency value in F_e, the self-correlation map $F_y(:, z, :, :)$ ($z = (x-1)W + y$ is the channel index) will be placed on the channel to generate the F_g.

III. Experiments

A. Experiment protocol

Training and test details. Our network is implemented based on Pytorch [31] with a single Nvdia V100 GPU. We use Adam [32] to optimize the proposed methods. The learning rate is 10^{-5}, and the weight decay is 10^{-4}. We choose a subset of the COCO dataset [33] as our training dataset, containing 9213 images as suggested by [25, 34–36]. The images are resized to 224×224 for training and testing. In our training, images are randomly flipped horizontally for augmentation, it takes about 3.5 hours to train our model with mini-batch size of 16 for 70 epochs. In the test phases, each image group with a random number of images constitutes a batch and regardless of its capacity, its co-saliency maps are generated at once.

Loss functions. In our network, IoU loss is used to supervise the predicted co-saliency maps M=$\{M_i\}_{i=1}^n (M_i \in \mathbb{R}^{H \times W})$ by the corresponding ground-truths G=$\{G_i\}_{i=1}^n (G_i \in \mathbb{R}^{H \times W})$ to well separate the foreground and background. It can be formulated as:

$$L(M, G) = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{j=1}^{HW} (min(M_i, G_i))_j}{\sum_{j=1}^{HW} (max(M_i, G_i))_j} \tag{6}$$

Evalution metrics and Datasets. We compare our ELR-Net with three challenging datasets for evaluation: Cosal2015 [37], CoSOD3k [20], and CoCA [38]. Cosal2015 is a large dataset widely used in the evaluation of CoSOD and it owns 2015 images of 50 categories, which is very challenging to accurately detect the co-salient regions. The last two were recently proposed for challenging real-world co-saliency evaluation. The CoSOD3k is the largest evaluation benchmark at present and it contains 3316 images of 160 categories. The CoCA contains 1297 images with 80 classes. Following the advice of recent work [20], we do not use iCoseg [39] and MSRC [40] for evaluation, because images in these datasets

TABLE I

TABLE I. PERFORMANCE COMPARISON WITH 9 STATE-OF-THE-ART METHODS ON 3 BENCHMARK DATASETS. MAXIMUM AND MEAN F-MEASURE (LARGER IS BETTER),MEAN E-MEASURE (LARGER IS BETTER), SMEASURE (SM, LARGER IS BETTER), AND MAE (SMALLER IS BETTER) ARE USED TO MEASURE THE MODEL PERFORMANCE. THE BEST RESULTS ARE HIGHLIGHTED IN RED. * REPRESENT SINGLE IMAGE SALIENCY OBJECT DETECTION METHODS.

Methods	CoSal2015					CoSOD3k					CoCA				
	Fm	mF	mE	Sm	MAE	Fm	mF	mE	Sm	MAE	Fm	mF	mE	Sm	MAE
F3Net* (AAAI2020)	0.818	0.809	0.863	0.842	0.084	0.747	0.737	0.818	0.794	0.100	0.440	0.427	0.623	0.613	0.178
MSFNet*(2021ACMM)	0.816	0.809	0.862	0.834	0.085	0.744	0.739	0.821	0.786	0.099	0.431	0.428	0.623	0.608	0.178
CBCD (TIP2013)	0.547	0.378	0.515	0.549	0.233	0.468	0.321	0.51	0.529	0.228	0.313	0.229	0.535	0.523	0.180
CSMG (CVPR2019)	0.787	0.721	0.763	0.776	0.130	0.730	0.596	0.675	0.727	0.141	0.498	0.390	0.606	0.627	0.114
ICNet(Neurips2020)	0.858	0.846	0.896	0.856	0.058	0.765	0.756	0.843	0.797	0.087	0.514	0.502	0.685	0.657	0.141
GCAGC(CVPR2020)	0.832	0.768	0.814	0.823	0.095	0.778	0.715	0.790	0.798	0.093	0.517	0.446	0.667	0.666	0.109
GICD(ECCV2020)	0.844	0.835	0.882	0.843	0.070	0.770	0.763	0.844	0.796	0.079	0.513	0.504	0.701	0.658	0.126
GCoNet(CVPR2021)	0.847	0.837	0.884	0.845	0.068	0.777	0.769	0.856	0.802	**0.071**	0.543	0.531	**0.739**	0.672	**0.105**
CoEGNet(TPAMI2021)	0.836	0.827	0.867	0.838	0.078	0.758	0.748	0.817	0.778	0.084	0.493	0.449	0.678	0.611	0.106
Ours	**0.869**	**0.858**	**0.903**	**0.859**	**0.057**	**0.785**	**0.777**	**0.857**	**0.808**	0.078	**0.547**	**0.536**	0.721	**0.677**	0.125

Fig. 6. Results of our ELR-Net compared with other state-of-the-art methods.

usually only have one salient object and are not very suitable for evaluation for CoSOD works. We use maximum F-measure F_m [41], mean F-measure mF, mean E-measure mE [42], S-measure S_m [43], and mean absolute error (MAE) [44] to evaluate methods in our experiments.

B. Comparison with State-of-the-arts

Comparison methods. Since not all CoSOD methods have publicly released codes. We compare our ELR-Net with seven state-of-the-art CoSOD methods and two cutting-edge deep salient object detection (SOD) methods. For CoSOD methods, we compare with one representative traditional algorithm (CBCD [10]) and six deep-based methods, including CSMG [18] , ICNet [25], GCAGC [19], GICD [38], GCoNet [22] and CoEGNet [20]. For SOD methods, we also compare our ELR-Net with F^3Net [45] and MSFNet [46].

Quantitative comparison. Table I shows the quantitative comparison results and shows ours ELR-Net achieves the best results on three challenges benchmarks by four widely used metrics. Figure 7 shows the PR curve and F-measure curve of all methods, our model is superior to other methods, where all the curves of ELR-Net are on the top of those generated by the other comparison methods.

Qualitative comparisons. To further prove the advantages

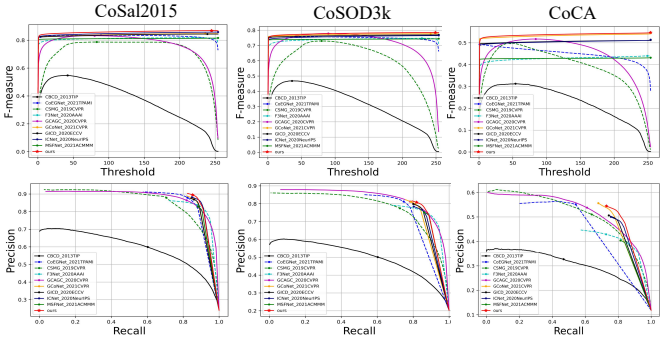

Fig. 7. Comparison with state-of-the-art methods in terms of PR and F-measure curves on three benchmark datasets.

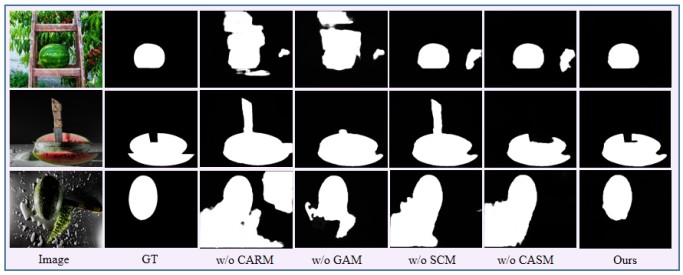

Fig. 8. **Qualitative ablation studies** of our ELR-Net on different modules.

of our model, Figure 6 shows the co-saliency maps generated by six representative state-of-the-art methods for further comparison.

C. Ablation Study

In this section, we study the effectiveness of each component in our model. Note that all variants have a similar capacity to ensure the performance gains are not due to the additional parameters.

Effectiveness of Co-attention and Refinement Module. In our network, CARM is designed to capture the common attribute of co-salient objects in an image group (Location) and further fine-grained information on the co-salient regions (Refinement). Here we validate the effectiveness of CARM for CoSOD. Table II lists the corresponding experimental results in terms of all datasets and metrics. For our model without "CARM" (i.e. w/o CARM), as reported in the 2^{nd} row of Table II, we can observe that the F_m score decreases by 7.4%, 8% and 12.4% on CoSal2015, CoSOD3k and CoCA, respectively. Moreover, without CRAM, other metrics of our model drop significantly. In summary, the CARM plays a vital role in our model by Location-Refinement mechanism. The results are visualized in 3^{rd} column of Figure 8

Effectiveness of GAM and SCM: We further research the importance of CARM, CARM contains GAM and SCM. GAM is designed to capture the common attributes of co-salient objects. We first remove the GAM (i.e. w/o GAM), the performance of our model drops significantly in terms of all metrics, as reported in the 3^{rd} row of Table II. Then, we test our model without SCM, which suffer from a drop

TABLE II
QUANTITATIVE ABLATION STUDIES OF OUR ELR-NET ON THE EFFECTIVENESS OF DIFFERENT MODULE.

Methods	CoSal2015		CoSOD3k		COCA	
	Fm	Sm	Fm	Sm	Fm	Sm
w/o CASM	0.855	0.850	0.768	0.795	0.513	0.644
w/o CARM	0.795	0.799	0.705	0.735	0.423	0.553
w/o GAM	0.850	0.843	0.754	0.784	0.508	0.644
w/o SCM	0.863	0.853	0.776	0.798	0.528	0.644
Ours	**0.869**	**0.859**	**0.785**	**0.808**	**0.547**	**0.677**

score of 0.6%, 1%, 3.3% in terms of S_m on all datasets, which proves the effectiveness of SCM to further refine the co-salient regions and boosts the result slightly due to the potential dependence on position is eliminated.The results are visualized in 4^{th} and 5^{th} columns of Figure 8.

Importance of CASM to our model In our network, CASM contains a channel and spatial attention module to enhance the high-level semantic features from the backbone network to obtain the co-features, which have common objects and highlight the location information of objects. To verify the effectiveness of CASM, we remove it from our network while keeping other modules unchanged (i.e., w/o CASM) and conduct the experiments on all datasets. The results are presented in the 1^{st} row of Table II. The results are visualized in 6^{th} column of Figure 8.

IV. CONCLUSION

In this paper, we proposed an Enhancement and Location-Refinement Network (ELR-Net) for co-saliency detection (CoSOD). Firstly, we propose a deep co-salient method based on channel and spatial attention module (CASM), which combines the attention mechanism to enhance the common semantic information. Secondly, we present a Co-attention and Refinement Module (CARM) to enhance the co-occurring salient regions by location and refinement. CARM contains GAM and SCM. GAM first enables the model to concentrate on the co-salient regions (Location). Then, SCM further fine-grained information on the co-salient regions (Refinement). Besides, single image saliency maps (SISMs) explore intra cues and further exploited correlations between intra cues and single image features to capture inter cues. Compared with 9 state-of-the-art methods on three public benchmarks datasets, our model achieves the best performance on quantitative and qualitative evaluation.

ACKNOWLEDGMENT

This research was funded by the National Natural Science Foundation of China (No.31870641), Special Funds of Science and Technology Innovation Project of Fujian Agriculture and Forestry University (No.KFA17030A and KFA17181A), the Natural Science Foundation of Fujian Province (No.2018J01612 and 2021J01125), and the Forestry Science and Technology Projects in Fujian Province, China (No. Memorandums 26).

REFERENCES

[1] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 186–202.

[2] M. Zhuge, D.-P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao, "Salient object detection via integrity learning," *arXiv preprint arXiv:2101.07663*, 2021.

[3] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7479–7489.

[4] W. Wang and J. Shen, "Higher-order image co-segmentation," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1011–1021, 2016.

[5] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Deepco3: Deep instance co-segmentation by co-peak search and co-saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8846–8855.

[6] Y. Zeng, Y. Zhuge, H. Lu, and L. Zhang, "Joint learning of saliency detection and weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7223–7233.

[7] K. R. Jerripothula, J. Cai, and J. Yuan, "Efficient video object co-localization with co-saliency activated tracklets," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 3, pp. 744–755, 2018.

[8] K. R. Jerripothula, J. Cai, and J. Yuan, "Cats: Co-saliency activated tracklet selection for video co-localization," in *European Conference on Computer Vision*. Springer, 2016, pp. 187–202.

[9] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in *CVPR 2011*. IEEE, 2011, pp. 2129–2136.

[10] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3766–3778, 2013.

[11] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3365–3375, 2011.

[12] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 4175–4186, 2014.

[13] K. Zhang, J. Chen, B. Liu, and Q. Liu, "Deep object co-segmentation via spatial-semantic network modulation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 813–12 820.

[14] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7264–7273.

[15] B. Jiang, X. Jiang, A. Zhou, J. Tang, and B. Luo, "A unified multiple graph learning and convolutional network model for co-saliency estimation," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1375–1382.

[16] C. Wang, Z.-J. Zha, D. Liu, and H. Xie, "Robust deep co-saliency detection with group semantic," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8917–8924.

[17] D. Zhang, D. Meng, C. Li, L. Jiang, Q. Zhao, and J. Han, "A self-paced multiple-instance learning framework for co-saliency detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 594–602.

[18] K. Zhang, T. Li, B. Liu, and Q. Liu, "Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3095–3104.

[19] K. Zhang, T. Li, S. Shen, B. Liu, J. Chen, and Q. Liu, "Adaptive graph convolutional network with attention graph clustering for co-saliency detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9050–9059.

[20] D.-P. Fan, T. Li, Z. Lin, G.-P. Ji, D. Zhang, M.-M. Cheng, H. Fu, and J. Shen, "Re-thinking co-salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[21] Q. Zhang, R. Cong, J. Hou, C. Li, and Y. Zhao, "Coadnet: Collaborative aggregation-and-distribution networks for co-salient object detection," *arXiv preprint arXiv:2011.04887*, 2020.

[22] Q. Fan, D.-P. Fan, H. Fu, C.-K. Tang, L. Shao, and Y.-W. Tai, "Group collaborative learning for co-salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 288–12 298.

[23] J. Chen, Y. Chen, W. Li, G. Ning, M. Tong, and A. Hilton, "Channel and spatial attention based deep object co-segmentation," *Knowledge-Based Systems*, vol. 211, p. 106550, 2021.

[24] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network," in *ECCV*, 2020.

[25] W.-D. Jin, J. Xu, M.-M. Cheng, Y. Zhang, and W. Guo, "Icnet: Intra-saliency correlation network for co-saliency detection," *Advances in Neural Information Processing Systems*, 2020.

[26] M. Siam, B. N. Oreshkin, and M. Jagersand, "Amp: Adaptive masked proxies for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5249–5258.

[27] C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, and K. Murphy, "Tracking emerges by colorizing

videos," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 391–408.
[28] Z. Lai and W. Xie, "Self-supervised learning for video correspondence flow," *arXiv preprint arXiv:1905.00875*, 2019.
[29] X. Wang, A. Jabri, and A. A. Efros, "Learning correspondence from the cycle-consistency of time," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2566–2576.
[30] Z. Lai, E. Lu, and W. Xie, "Mast: A memory-augmented self-supervised tracker," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6479–6488.
[31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.
[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
[34] B. Li, Z. Sun, L. Tang, Y. Sun, and J. Shi, "Detecting robust co-saliency with recurrent co-attention neural network." in *IJCAI*, vol. 2, 2019, p. 6.
[35] L. Wei, S. Zhao, O. E. F. Bourahla, X. Li, and F. Wu, "Group-wise deep co-saliency detection," *arXiv preprint arXiv:1707.07381*, 2017.
[36] X. Zheng, Z.-J. Zha, and L. Zhuang, "A feature-adaptive semi-supervised framework for co-saliency detection," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 959–966.
[37] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *International Journal of Computer Vision*, vol. 120, no. 2, pp. 215–232, 2016.
[38] Z. Zhang, W. Jin, J. Xu, and M.-M. Cheng, "Gradient-induced co-saliency detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 455–472.
[39] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "icoseg: Interactive co-segmentation with intelligent scribble guidance," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3169–3176.
[40] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2. IEEE, 2005, pp. 1800–1807.
[41] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 1597–1604.
[42] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 2018, pp. 698–704.
[43] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4548–4557.
[44] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *Proceedings of the IEEE International Conference on Computer vision*, 2013, pp. 1529–1536.
[45] J. Wei, S. Wang, and Q. Huang, "F^3net: Fusion, feedback and focus for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 321–12 328.
[46] M. Zhang, T. Liu, Y. Piao, S. Yao, and H. Lu, "Auto-msfnet: Search multi-scale fusion network for salient object detection," in *ACM Multimedia Conference 2021*, 2021.