Re-thinking Co-Salient Object Detection

Deng-Ping Fan, Tengpeng Li, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Ming-Ming Cheng, *Senior Member, IEEE*, Huazhu Fu, *Senior Member, IEEE*, Jianbing Shen, *Senior Member, IEEE*

Abstract—In this paper, we conduct a comprehensive study on the co-salient object detection (CoSOD) problem for images. CoSOD is an emerging and rapidly growing extension of salient object detection (SOD), which aims to detect the co-occurring salient objects in a group of images. However, existing CoSOD datasets often have a serious data bias, assuming that each group of images contains salient objects of similar visual appearances. This bias can lead to the ideal settings and effectiveness of models trained on existing datasets, being impaired in real-life situations, where similarities are usually semantic or conceptual. To tackle this issue, we first introduce a new benchmark, called CoSOD3k in the wild, which requires a large amount of semantic context, making it more challenging than existing CoSOD datasets. Our CoSOD3k consists of 3,316 high-quality, elaborately selected images divided into 160 groups with hierarchical annotations. The images span a wide range of categories, shapes, object sizes, and backgrounds. Second, we integrate the existing SOD techniques to build a unified, trainable CoSOD framework, which is long overdue in this field. Specifically, we propose a novel CoEG-Net that augments our prior model EGNet with a co-attention projection strategy to enable fast common information learning. CoEG-Net fully leverages previous large-scale SOD datasets and significantly improves the model scalability and stability. Third, we comprehensively summarize 40 cutting-edge algorithms, benchmarking 18 of them over three challenging CoSOD datasets (iCoSeg, CoSal2015, and our CoSOD3k), and reporting more detailed (*i.e.*, group-level) performance analysis. Finally, we discuss the challenges and future works of CoSOD. We hope that our study will give a strong boost to growth in the CoSOD community. The benchmark toolbox and results are available on our project page at http://dpfan.net/CoSOD3K/.

Index Terms—Co-saliency Detection, Co-attention Projection, CoSOD Dataset, Benchmark.

1 Introduction

Salient object detection (SOD) in color images [2]–[6], RGB-D images [7]–[11], and videos [12]–[14] has been an active field of research in the computer vision community over the past [15]–[22]. SOD mimics the human vision system to detect the most attention-grabbing object(s) in a single image, as shown in Fig. 1 (a). As a extension of this, co-salient object detection (CoSOD) emerged recently to employ a set of images. The goal of CoSOD is to extract the salient object(s) that are common within a single image (*e.g.*, red-clothed football players in Fig. 1 (b)) or across multiple images (*e.g.*, the blue-clothed gymnast in Fig. 1 (c)). Two important characteristics of co-salient objects are local saliency and global similarity. Due to its useful potential, CoSOD has been attracting growing attention in many applications, including collection-aware crops [23], co-segmentation [24], [25], weakly supervised learning [26], image retrieval [27], [28], and video foreground detection [29].

As such, the CoSOD task has been rapidly growing in recent few years [18], [34], with hundreds of related publications since 2010[1]. Most CoSOD datasets tend to focus on the appearance-similarity between objects to identify the co-salient object across multiple images. However, this leads to *data selection bias* [2], [35] and is not always appropriate, since, in real-world applications, the salient objects in a group of images often vary in terms of *texture*, *scene*, and *background* (see our CoSOD3k dataset in Fig. 1 (d)), even if they belong to the same category. In addition to the data selection bias, CoSOD methods also suffer from two main limitations:

(A) Completeness. ϵ (Mean Absolute Error) [36] and F-measure [37] are two widely used metrics in CoSOD/SOD model evaluation. As discussed in [38], these metrics have their inherent limitations. To provide thorough and reliable conclusions, we need introduce more accurate metrics *e.g.*, structural based evaluation metric or perceptual based evaluation metric.

(B) Fairness. To evaluate the F-measure, the first step is to binarize a saliency map into a set of foreground maps using different threshold values. There are many binarization strategies [39], such as adaptive threshold, fixed threshold and so on. However, different strategies will result in different F-measure performances. Further, few previous works provide details on their binarization strategy, leading to inconsistent F-measures for different researchers.

To address the aforementioned limitations, we argue that integrating various publicly available CoSOD algorithms, datasets, and metrics, and then providing a complete, unified benchmark, is highly desired. As such, we make four distinct contributions in this work:

- *D.-P. Fan, Z. Lin and M.-M. Cheng are with the College of Computer Science, Nankai University, Tianjin, China. (Email: dengpfan@gmail.com, frazer.linzheng@gmail.com, cmm@nankai.edu.cn)*
- *T. Li is with the B-DAT and CICAEET, Nanjing University of Information Science and Technology, Nanjing, China. (E-mail: ltpfor1225@gmail.com)*
- *G.-P. Ji is with the School of Computer Science, Wuhan University, Hubei, China. (E-mail: gepengai.ji@gmail.com)*
- *D. Zhang is with the Brain and Artificial Intelligence Laboratory, School of Automation, Northwestern Polytechnical University, Xi'an 710072, China. (E-mail: zhangdingwen2006yyy@gmail.com)*
- *H. Fu and J. Shen are with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. (E-mail: {huazhu.fu, jianbing.shen}@inceptioniai.org)*
- *A preliminary version of this work has appeared in CVPR 2020 [1].*
- *Corresponding author: M.-M. Cheng.*

1. Some representative works can be found on https://hzfu.github.io/proj_cosal_review.html.

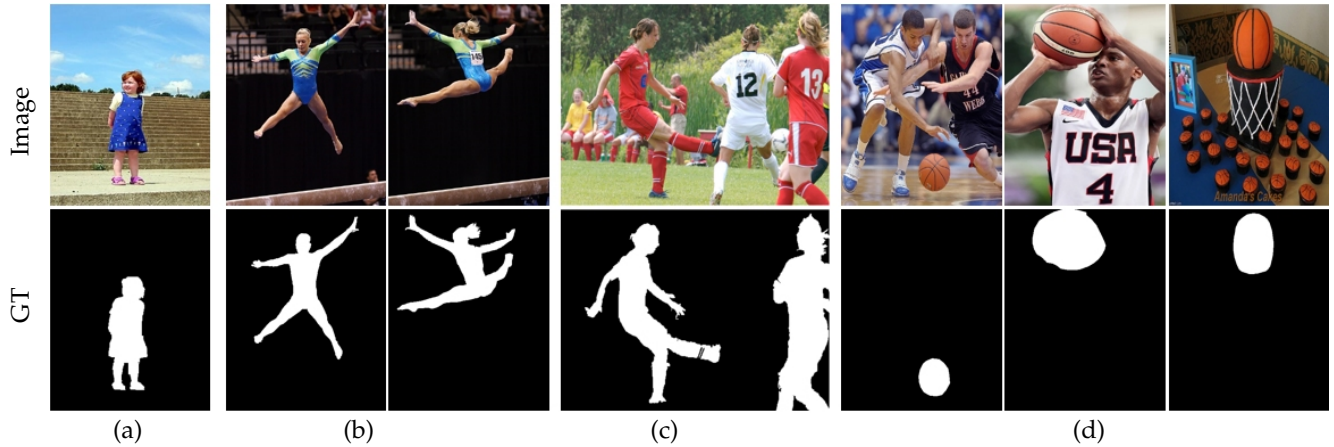


Fig. 1. Different salient object detection (SOD) tasks. (a) Traditional SOD [30]. (b) Within-image co-salient object detection (CoSOD) [31], where common salient objects are detected from a single image. (c) Existing CoSOD, where salient objects are detected across a pair [32] or a group [33] of images with similar appearances. (d) The proposed CoSOD in the wild, which requires a large amount of semantic context, making it more challenging than existing CoSOD.

- **First, we construct a challenging CoSOD3k dataset, with more realistic settings.** Our CoSOD3k[2] is the largest CoSOD dataset to date, with two aspects: 1) it contains 13 super-classes, 160 groups, and 3,316 images in total, where each super-class is carefully selected to cover diverse scenes; *e.g.*, *Vehicle*, *Food*, *Tool*, *etc.*; 2) each image is accompanied by hierarchical annotations, including category, bounding box, object, and instance, which could greatly benefit various vision tasks (*e.g.*, object proposal, co-location, co-segmentation, co-instance detection, *etc.*), as shown in Fig. 2.
- **Second, we present the first large-scale co-salient object detection study,** reviewing 40 state-of-the-art (SOTA) models, and evaluating 18 of them on three challenging, large-scale CoSOD datasets (iCoSeg, CoSal2015, and the proposed CoSOD3k). A convenient benchmark toolbox is also provided to integrate various publicly available CoSOD datasets and multiple metrics for better performance evaluation. The benchmark toolbox and results have been made publicly available at https://dpfan.net/CoSOD3K/.
- **Third, we propose a simple but effective *CoEG-Net* baseline for CoSOD,** which uniformly and simultaneously embeds the appearance and semantic features through a co-attention projection and a basic SOD network. Comprehensive benchmarking results show that *CoEG-Net* outperforms the 18 SOTA models. Moreover, it also yields competitive visual results, making it an efficient solution for the CoSOD task.
- **Finally, we make several interesting observations, discuss the important issues arising from the benchmark results, and suggest some future directions.** Our study serves as a potential catalyst for promoting large-scale model comparison for future CoSOD research.

2. Collecting the CoSOD dataset is more difficult than the SOD dataset, that is why the previous largest CoSOD dataset, *i.e.*, [40], in the past 15 years has only 2K images. Even for our 3K dataset, we have spent 1 year to collect such high-quality dataset. Moreover, we also pay more attention to provide high-quality hierarchical annotations (*e.g.*, image-level and object-/instance-level) to promote related vision tasks rather than the size of the dataset.

TABLE 1
Statistics of existing CoSOD datasets and the proposed CoSOD3k, showing that CoSOD3k provides higher-quality and much richer annotations. **#Gp**: number of image groups. **#Img**: number of images. **#Avg**: average number of images per group. **IL**: whether or not instance-level annotations are provided. **Ceg**: whether or not category labels are provided for each group. **BBx**: whether or not bounding box labels are provided for each image. **HQ**: high-quality annotation.

Dataset	Year	#Gp	#Img	#Avg	IL	Ceg	BBx	HQ	Input
MSRC [33]	2005	8	240	30					Group images
iCoSeg [41]	2010	38	643	17				✓	Group images
Image Pair [32]	2011	105	210	2		✓*			Two images
CoSal2015 [40]	2015	50	2,015	40		✓*		✓	Group images
WICOS [31]	2018	364	364	1				✓	Single image
CoSOD3k	2020	160	3,316	21	✓	✓	✓	✓	Group images

* denotes coarse category rather than explicitly accurate category.

This paper is based on and extends our previous CVPR2020 version [1] in the following aspects. 1) We have implemented a simple but effective framework of CoSOD, which uniformly and simultaneously embeds the appearance and semantic features through a sparse convolution and a basic SOD network. Importantly, we also designed a common feature detector, which solved with Plug-and-Play. 2) We have made a lot of efforts to improve the presentations (*e.g.*, dataset, framework, key results) and organizations of our paper. We have added several new sections to describe our new framework about the method formulation, corresponding technical components, and further experiments (*e.g.*, comparison with baselines, running time). Besides, several sections have been re-written to improve the readability and provide more detailed explanations about the introduction, CoSOD models, quantitative/qualitative comparisons, and discussions. 3) We build the first standard Benchmark and model zoo of CoSOD, which integrates various publicly available CoSOD datasets with uniform input/output formats (*i.e.*, JPEG for image; PNG for GT). The gathered code of traditional or learning-based will be released soon as well.

TABLE 2
Summary of 40 classic and cutting-edge CoSOD approaches. **Training set:** PV = PASCAL VOC07 [42]. CR = Coseg-Rep [43]. DO = DUT-OMRON [44]. COS = COCO-subset. **Main Component:** IMC = Intra-Image Contrast. IGS: Intra-Group Separability. IGC: Intra-Group Consistency. SPL: Self-Paced Learning. CH: Color Histogram. GMR: Graph-based Manifold Ranking. CAE: Convolutional Auto Encoder. HSR: High-spatial Resolution. FSM: five saliency models including CBCS [29], RC [45], DCL [16], RFCN [46], DWSI [31]. **SL.** = Supervision Level. W = Weakly-supervised. S = Supervised. U = Unsupervised. **Sp.:** Whether or not superpixel techniques are used. **Po.:** Whether or not proposal algorithms are utilized. **Ed.:** Whether or not edge features are explicitly used. **Post.:** Whether or not post-processing methods, such as, CRF [47], GraphCut (GCut), or adaptive/constant threshold (THR), are introduced. ‡ denotes deep models. More details about these models can be found in recent survey papers [1], [18], [34].

#	Model	Pub. Year	#Training	Training Set	Main Component	SL.	Sp.	Po.	Ed.	Post.
1	WPL [23]	UIST 2010			Morphological, Translational Alignment	U				
2	PCSD [48]	ICIP 2010	120,000	8*8 image patch	Sparse Feature [49], Filter Bank	W				
3	IPCS [32]	TIP 2011			Ncut, Co-multilayer Graph	U	✓			
4	CBCS [29]	TIP 2013			Contrast/Spatial/Corresponding Cue	U				
5	MI [50]	TMM 2013			Feature/Images Pyramid, Multi-scale Voting	U	✓			GCut
6	CSHS [51]	SPL 2013			Hierarchical Segmentation, Contour Map [52]	U			✓	
7	ESMG [53]	SPL 2014			Efficient Manifold Ranking [54], OTSU [55]	U				
8	BR [56]	MM 2014			Common/Center Cue, Global Correspondence	U	✓			
9	SACS [57]	TIP 2014			Self-adaptive Weight, Low Rank Matrix	U	✓			
10	DIM‡ [58]	TNNLS 2015	1,000 + 9,963	ASD [37] + PV	SDAE Model [58], Contrast/Object Prior	S	✓			
11	CODW‡ [59]	IJCV 2016		ImageNet [60] pre-train	SermaNet [61], RBM [62], IMC, IGS, IGC	W	✓	✓		
12	SP-MIL‡ [63]	TPAMI 2017	(240+643)*10%	MSRC-V1 [33] + iCoSeg [41]	SPL [64], SVM, GIST [65], CNNs [66]	W	✓			
13	GD‡ [67]	IJCAI 2017	9,213	MSCOCO [68]	VGGNet16 [69], Group-wise Feature	S				
14	MVSRCC‡ [70]	TIP 2017			LBP, SIFT [71], CH, Bipartite Graph		✓	✓		
15	UMLF [72]	TCSVT 2017	(240 + 2015)*50%	MSRC-V1 [33] + CoSal2015 [59]	SVM, GMR [44], Metric Learning	S	✓			
16	DML‡ [73]	BMVC 2018	10,000 + 6,232 + 5,168	M10K [45] + THUR15K [28] + DO	CAE, HSR, Multistage	S				
17	DWSI [31]	AAAI 2018			EdgeBox [74], Low-rank Matrix, CH	S		✓		
18	GONet‡ [75]	ECCV 2018		ImageNet [60] pre-train	ResNet-50 [76], Graphical Optimization	W	✓			CRF
19	COC‡ [77]	IJCAI 2018		ImageNet [60] pre-train	ResNet-50 [76], Co-attention Loss	W		✓		CRF
20	FASS‡ [78]	MM 2018		ImageNet [60] pre-train	DHS [79]/VGGNet, Graph Optimization	W	✓			
21	PJO [80]	TIP 2018			Energy Minimization, BoWs	U	✓			
22	SPIG‡ [81]	TIP 2018	10,000+210 +2015+240	M10K [45]+IPCS [32] + CoSal2015 [59] + MSRC-V1 [33]	DeepLab, Graph Representation	S	✓			
23	QGF [82]	TMM 2018		ImageNet [60] pre-train	Dense Correspondence, Quality Measure	S	✓			THR
24	EHL‡ [83]	NC 2019	643	iCoSeg [41]	GoogLeNet [84], FSM	S	✓			
25	IML‡ [85]	NC 2019	3624	CoSal2015 [59] + PV + CR	VGGNet16 [69]	S	✓			
26	DGFC‡ [86]	TIP 2019	>200,000	MSCOCO [68]	VGGNet16 [69], Group-wise Feature	S	✓			
27	RCANet‡ [87]	IJCAI 2019	>200,000	MSCOCO [68] + COS + iCoSeg [41] + CoSal2015 [59] + MSRC [33]	VGGNet16 [69], Recurrent Units	S				THR
28	GS‡ [88]	AAAI 2019	200,000	COCO-SEG [88]	VGGNet19 [69], Co-category Classification	S				
29	MGCNet‡ [89]	ICME 2019			Graph Convolutional Networks [90]	S	✓			
30	MGLCN‡ [91]	MM 2019	N/A	N/A	VGGNet16, PiCANet [92], Inter-/Intra-graph	S	✓			
31	HC‡ [93]	MM 2019	N/A	N/A	VAE-Net [94], Hierarchical Consistency	S	✓	✓		CRF
32	CSMG‡ [95]	CVPR 2019	25,00	MB [96]	VGGNet16 [69], Shared Superpixel Feature	S	✓			
33	DeepCO^3‡ [97]	CVPR 2019	10,000	M10K [45]	SVFSal [98] / VGGNet [69], Co-peak Search	W		✓		
34	GWD‡ [99]	ICCV 2019	>200,000	MSCOCO [68]	VGGNet19 [69], RNN, Group-wise Loss	S				THR
35	CAFCN‡ [100]	TCSVT 2020	200,000	MSCOCO [68]	VGGNet16 [69], Co-Attention, FCN	S				
36	GSPA‡ [101]	TNNLS 2020	200,000	COCO-SEG [101]	VGGNet19 [69], Group Semantic, Pyramid Attention	S				
37	GOMAG [102]	TMM 2020	N/A	N/A	General Optimization, Adaptive Graph Learning	U	✓			
38	AGC‡ [103]	CVPR 2020	200,000	MSCOCO [68]	VGGNet16 [69], Graph Convolution & Clustering	S				
39	GICD‡ [104]	ECCV 2020	8,250	DUTS [30]	VGGNet19 [69], Gradient Inducing, Attention Retaining	S				
40	CoEG-Net‡ (Ours)	2020	10,553	DUTS [30]	VGGNet16, Co-attention Projection	S			✓	CRF

2 Related Work

2.1 CoSOD Datasets

Currently, only a few CoSOD datasets have been proposed [28], [31]–[33], [40], [41], as shown in Table 1. *MSRC* [33] and *Image Pair* [32] are two of the earliest ones. *MSRC* was designed for recognizing object classes from images and has spurred many interesting ideas over the past several years. This dataset includes 8 image groups and 240 images in total, with manually annotated pixel-level ground-truth data. *Image Pair*, introduced by Li *et al.* [32], was specifically designed for image pairs and contains 210 images (105 groups) in total. The *iCoSeg* [41] dataset was released in 2010. It is a relatively larger dataset consisting of 38 categories with 643 images in total. Each image group in this dataset contains 4 to 42 images, rather than only 2 images like in the *Image Pair* dataset. The *THUR15K* [28] and *CoSal2015* [40] are two large-scale publicly available datasets, with CoSal2015 widely used for assessing co-salient object detection algorithms. Different from the above-mentioned datasets, the *WICOS* [31] dataset aims to detect co-salient objects from a single image, where each image can be viewed as one group.

Although the aforementioned datasets have advanced the CoSOD task to various degrees, they are severely limited in variety, with only dozens of groups. On such small-scale datasets, the scalability of methods cannot be fully evaluated. Moreover, these datasets only provide object-level labels. None of them provide rich annotations such as bounding boxes, instances, *etc.*, which are important for progressing many vision tasks and multi-task modeling. Especially in the current deep learning era, where models are often data-hungry. In this work, thus, we will focus on the two relatively large-scale datasets (*i.e.*, *iCoSeg* [41] and *CoSal2015* [40]) together with the proposed challenging dataset to provide more in-depth analysis.

Fig. 2. Sample images from our CoSOD3k dataset. It has rich annotations, *i.e.*, image-level categories (top), bounding boxes, object-level masks, and instance-level masks. Our CoSOD3k will provide a solid foundation for the CoSOD task and can benefit a wide range of related fields, *e.g.*, co-segmentation, weakly supervised localization.

2.2 CoSOD Methods.

Previous CoSOD studies [32], [57], [72], [80] have found that the inter-image correspondence can be effectively modeled by segmenting the input image into several computational units (*e.g.*, superpixel regions [105], or pixel clusters [29]). A similar observation can be found in recent reviews [18], [34]. In these approaches, heuristic characteristics (*e.g.*, contour [51], color, luminance) are extracted from images, and the high-level features are captured to express the semantic attributes in different ways, such as through metric learning [72] or self-adaptive weighting [57]. Several studies have also investigated how to capture inter-image constraints through various computational mechanisms, such as translational alignment [23], efficient manifold ranking [53], and global correspondence [56]. Some methods (*e.g.*, PCSD [48], which only uses a filter bank technique) do not even need to perform the correspondence matching between the two input images, and are able to achieve CoSOD before the co-attention occurs.

Recently, deep learning based CoSOD models have achieved good performance by learning co-salient object representations jointly. For instance, Zhang *et al.* [58] introduced a domain adaption model to transfer prior knowledge for CoSOD. Wei *et al.* [67] used a group input and output to discover the collaborative and interactive relationships between group-wise and single-image feature representations, in a collaborative learning framework. Along another line, the MVSRCC [70] model employs typical features, such as SIFT, LBP, and color histograms, as multi-view features. In addition, several other methods [77], [81], [83], [86], [88], [95], [97] are based on more powerful CNN models (*e.g.*, ResNet [76], Res2Net [106], GoogLeNet [84], and VGGNet [69]), achieving SOTA performances. These deep models generally achieve better performance through either weakly-supervised (*e.g.*, CODW [59], SP-MIL [63], GONet [75], and FASS [78]) or fully supervised learning (*e.g.*, DIM [58] and GD [67], and DML [73]). There are also some concurrent works [100]–[104] that are proposed after this submission. A summary of the existing CoSOD models is provided in Table 2.

3 CoSOD3k Dataset

3.1 Image Collection

We build a high-quality dataset, CoSOD3k, images of which are collected from the large-scale object recognition dataset ILSVRC [107]. There are several benefits of using ILSVRC to generate our dataset. First, ILSVRC is gathered from *Flickr* using scene-level queries and thus it includes various object categories, diverse realistic-scenes, and different object appearances, and covers a large span of the major challenges in CoSOD, providing us a solid basis for building a representative benchmark dataset for CoSOD. More importantly, though, the accompanying axis-aligned bounding boxes for each target object category allow us to identify unambiguous instance-level annotations.

3.2 Hierarchical Annotation

Similar to [108], [109], the data annotation is performed in a hierarchical (coarse to fine) manner (see Fig. 2).

• **Category Labeling.** We establish a hierarchical (three-level) taxonomic system for the CoSOD3k dataset. 160 common categories (see Fig. 3) are selected to generate *sub-classes* (*e.g.*, *Ant*, *Fig*, *Violin*, *Train*, *etc.*), which are consistent with the original categories in ILSVRC. Then, an upper-level class (*middle-level*) is assigned for each *sub-class*. Finally, we integrate the upper-level classes into 13 *super-classes*. The taxonomic structure of our CoSOD3k is given in Fig. 4.

• **Bounding Box Labeling.** The second level of annotation is bounding box labeling, which is widely used in object detection and localization. Although the ILSVRC dataset provides bounding box annotations, the labeled objects are not necessarily salient. Following many famous SOD datasets [30], [37], [45], [96], [110]–[116], we ask three viewers to redraw the bounding boxes around the object(s) in each image that dominate their attention. Then, we merge the bounding boxes labeled by the three viewers and have two additional senior researchers in the CoSOD field double-check the annotations. After that, as done in [117], we discard the images that contain more than six objects. Finally, we collect 3,316 images within 160 categories. Examples can be found in Fig. 2.

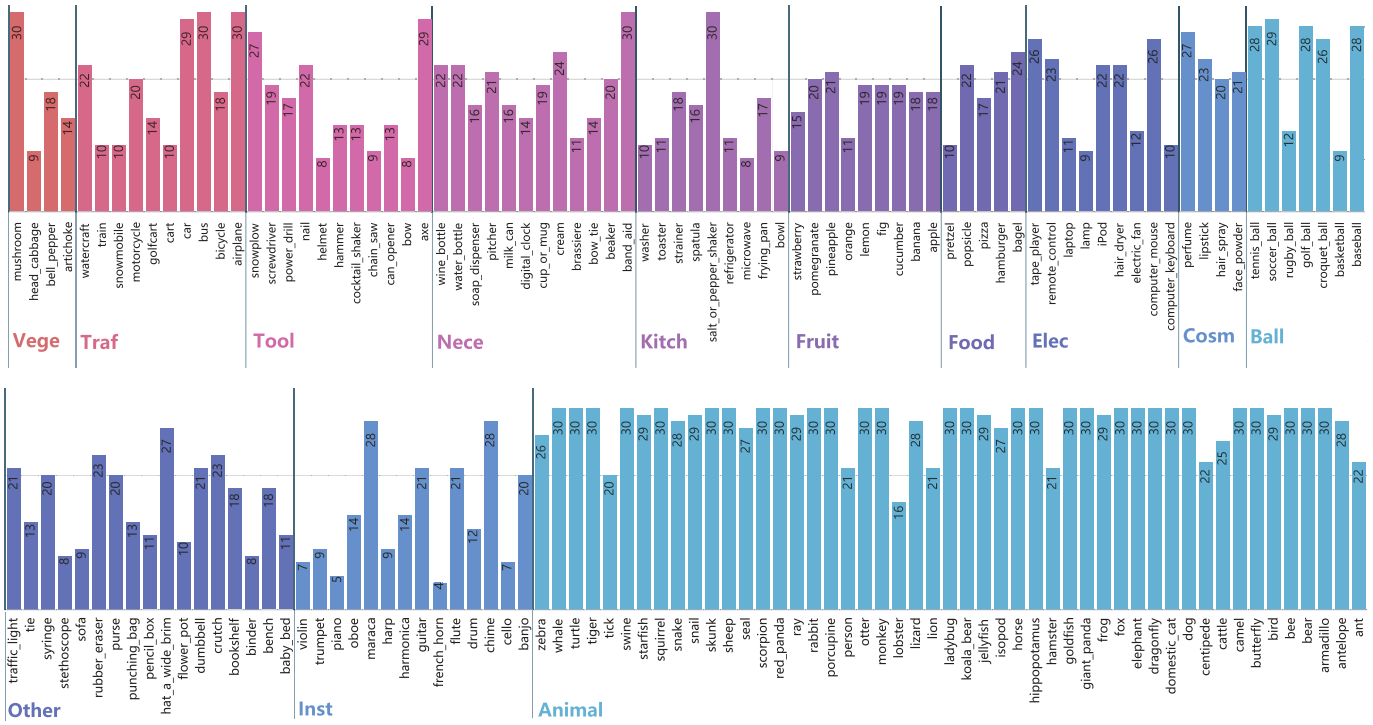


Fig. 3. Number of images in the 160 sub-classes of our dataset. Best viewed on screen and zoomed-in for details.

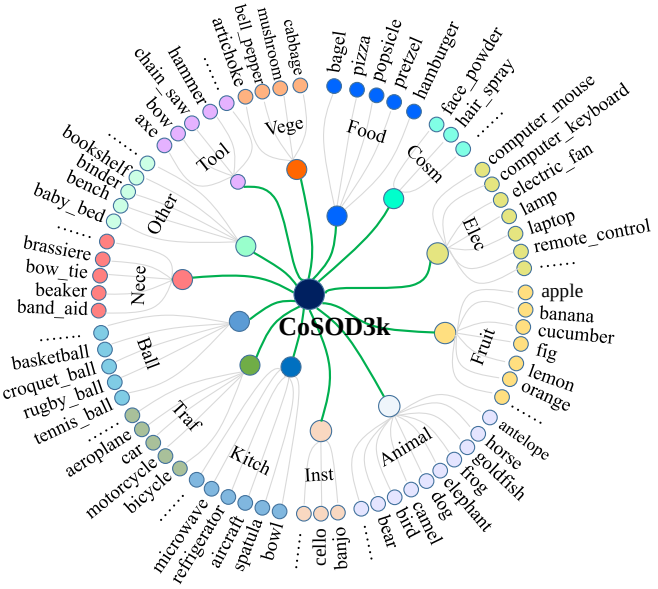


Fig. 4. Taxonomic structure of our dataset, which contains 13 super-classes with 160 sub-classes.

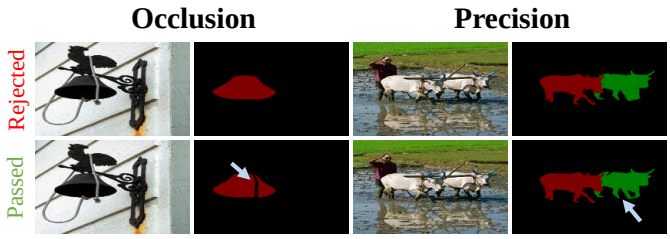


Fig. 5. Some passed and rejected cases (*e.g.*, occlusion, precision) in our CoSOD3k.

• **Object-/Instance-level Annotation.** High-quality pixel-level masks are necessary for CoSOD datasets. We hire twenty professional annotators and train them with 100 image examples. They are then instructed to annotate the images with object- and instance-level labels according to the previous bounding boxes. The average annotation time per image is about 8 and 15 minutes for object-level and instance-level labeling, respectively. Moreover, we also have three volunteers cross-check the whole process (more than three-fold), to ensure high-quality annotation (see Fig. 5). In this way, we obtain an accurate and challenging dataset with a total of 3,316 object-level and 4,915 instance-level annotations. Note that our final bounding box labels are refined further based on the instance-level annotations to tighten the target.

3.3 Dataset Features and Statistics

To provide deeper insight into our CoSOD3k, we present several important characteristics below.

• **Mixture-specific Category Masks.** Fig. 7 shows the average ground-truth masks for individual categories and the overall dataset. As can be observed, some categories with unique shapes (*e.g.*, airplane, zebra, and bicycle) present shape-biased maps, while categories with non-rigid or convex shapes (*e.g.*, goldfish, bird, and bus) do not have clear shape-bias. The overall dataset mask (the right of Fig. 7) tends to appear as a center-biased map without shape bias. As is well-known, humans are usually inclined to pay more attention to the center of a scene when taking a photo. Thus, it is easy for a SOD model to achieve a high score when employing a Gaussian function in its algorithm. Due to the limitation of space, we present all 160 mixture-specific category masks in the supplementary materials.

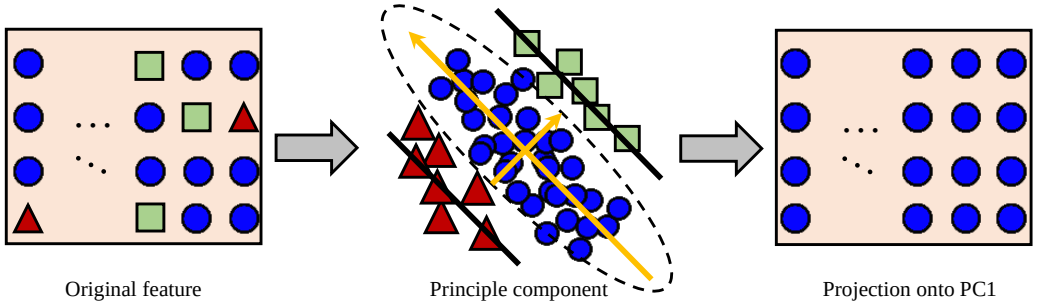


Fig. 6. Illustration of our co-attention projection operation. Given the original feature representation which covers common objects (circle), noisy foregrounds (triangle) and background clutter (square), the co-attention projection identifies the principle components of common objects, helping to preserve the common objects while removing interference. By adopting our co-attention projection operation, we finally project the principle component and obtain the new feature representation. Please refer to Section 4.2 for more details.

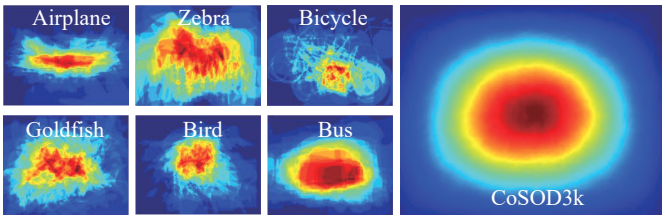


Fig. 7. Visualization of overlap masks for mixture-specific category and overall dataset masks of our CoSOD3k.

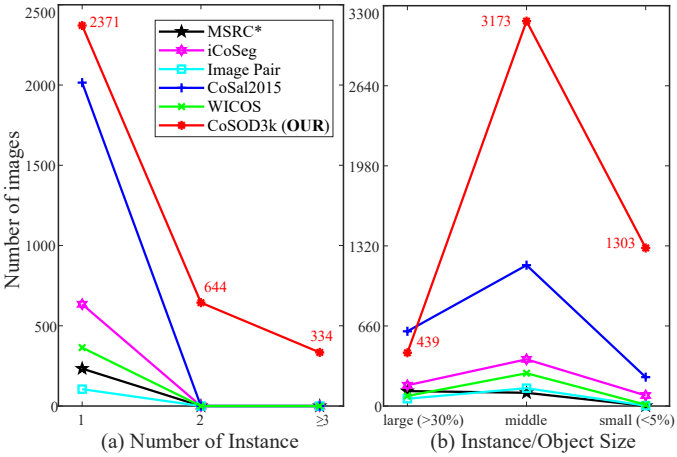


Fig. 8. The number of images for the MSRC, iCoSeg, Image Pair, CoSal2015, WICOS, and our CoSOD3k dataset in terms of the number of instances (a) and the instance/object size (b).

• **Sufficient Object Diversity.** As shown in Table 5 (2^{nd} row) and Fig. 3, our CoSOD3k covers a large variety of super-classes including *Vegetables*, *Food*, *Fruit*, *Tool*, *Necessary*, *Traffic*, *Cosmetic*, *Ball*, *Instrument*, *Kitchenware*, *Animal*, and *Others*, enabling a comprehensive understanding of real-world scenes.

• **Number of Instances.** Being able to parse objects into instances is critical for humans to understand, categorize, and interact with the world. To enable learning methods to gain instance-level understanding, annotations with instance labels are in high demand. With this in mind, in contrast to existing CoSOD datasets, our CoSOD3k contains the multi-instance scenes with instance-level annotations. As illustrated in Fig. 8 (a), the number of instances (1, 2, $\geq$3) is subject to a ratio of 7:2:1.

• **Size of Instances.** The instance size is defined as the ratio of foreground instance pixels to the total image pixels. Fig. 8 (b) shown the instance sizes of our CoSOD3k in terms of small, middle, and large instance/object. The distributions of instance sizes are $0.02\% \sim 86.5\%$ (avg.: 13.8%), yielding a broad range.

4 Proposed Method

In this work, we also propose a simple but effective *CoEGNet* baseline for CoSOD, which extend state-of-the-art SOD model EGNet [118] by introducing co-attention information in an unsupervised manner.

4.1 Method Formulation

For a group of N associated images $\{\mathbf{I}^n\}_{n=1}^N$, the co-saliency detection task aims at segmenting out the common attentive foreground objects and generating optimized co-saliency maps, which indicate common salient objects among the input images. To predict the co-saliency masks, we present a two-branch detection framework to respectively capture the concurrent dependencies and salient foregrounds in a multiply independent fashion. Fig. 9 illustrates the framework of the proposed method, which independently outputs co-attention maps $\{\mathbf{A}^n\}_{n=1}^N$ in the top branch and saliency prior maps $\{\mathbf{S}^n\}_{n=1}^N$ in the bottom branch. The co-attention map $\mathbf{A}^n$ and saliency prior map $\mathbf{S}^n$ are then integrated via element-wise multiply to produce the final co-saliency prediction $\mathbf{A}^n \otimes \mathbf{S}^n$.

To obtain the saliency prior map $\mathbf{S}^n$ for an input image $\mathbf{I}^n$, we simply use the edge guided salient object detection method EGNet [118] to collect multi-scale saliency priors. The EGNet is trained on large scale single image SOD dataset DUTS [30], which helps to identify the salient object regions in images without cross image information. The real challenge then becomes how to discover co-attention map $\mathbf{A}^n$ in an unsupervised manner, which we present in the next subsection.

4.2 Co-attention Projection for Co-saliency Learning

The design of co-attention learning (see Fig. 6) is motivated by the class activation mapping (CAM) technique proposed by Zhou *et al.* [119]. Given an input image $\mathbf{I}^n$, the corresponding feature **activations** $\mathbf{X}^n$ in the last convolution

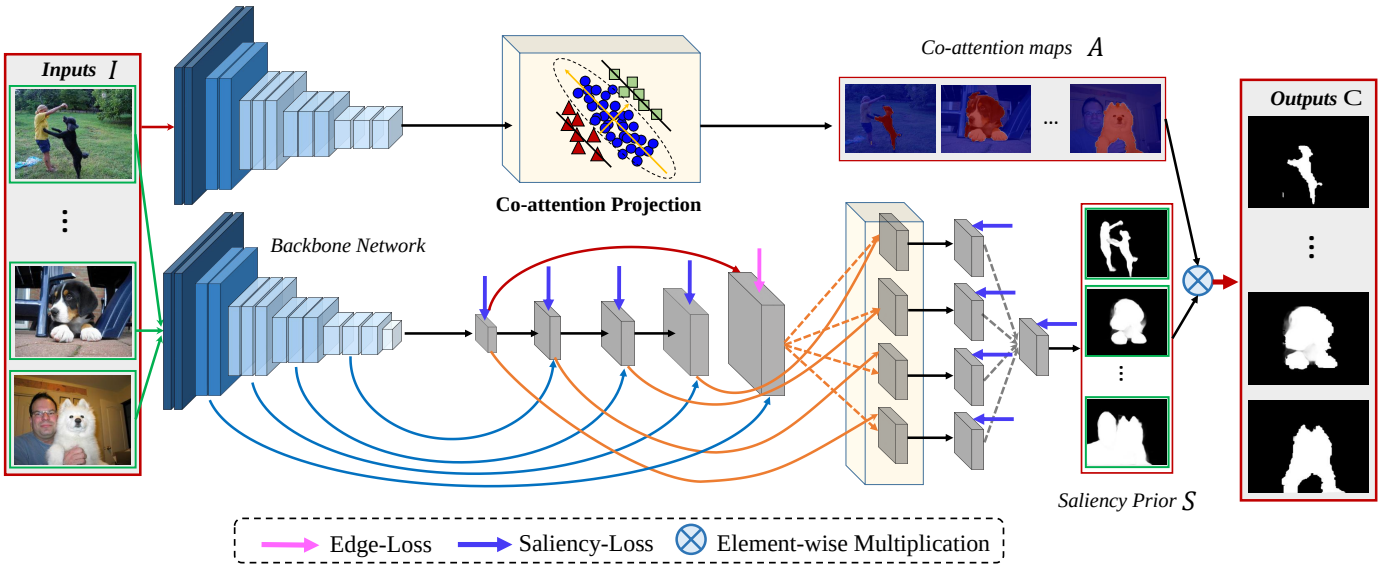


Fig. 9. Pipeline of the proposed architecture which contains two separate branches. For a group of images $\{\mathbf{I}^n\}_{n=1}^N$ as inputs, in the top branch, the extracted high-level image features are fed into the co-attention projection module to produce a co-attention map $\mathbf{A}^n$ for each input image $\mathbf{I}^n$. In the bottom branch, each image $\mathbf{I}^n$ is sent into the edge-guided saliency detection network (EGNet) [118] to generate the saliency prior map $\mathbf{S}^n$. Finally, $\mathbf{A}^n$ and $\mathbf{S}^n$ are simply integrated using element-wise multiply to produce the optimized outputs $\mathbf{A}^n \otimes \mathbf{S}^n$. See Section 4 for details.

TABLE 3
Table of symbols, their dimensions, indices, and meaning.

Symbol	Dimensions	Indices	Meaning
$\mathbf{A}^n$	$H \times W$	(i,j)	co-attention map of $\mathbf{I}^n$
$\mathbf{S}^n$	$H \times W$	(i,j)	saliency prior map of $\mathbf{I}^n$
$\mathbf{X}^n$	$H \times W \times K$	(i,j,k)	**activations** of the last conv layer
$\mathbf{X}_k^n$	$H \times W$	(i,j)	**feature map** of a channel in $\mathbf{X}^n$
H	1×1	scalar	spatial height
W	1×1	scalar	spatial width
K	1×1	scalar	number of feature channels
$\mathbf{x}^n(i,j)$	$K \times 1$	k	**descriptor** of $\mathbf{X}^n$ at location (i,j)
$\mathbf{M}_c^n$	$H \times W$	(i,j)	attention map for class c
ω^c	$K \times 1$	k	channel-wise weights for class c
$\bar{\mathbf{x}}$	$K \times 1$	k	average value of all $\mathbf{x}^n(i,j)$
$\hat{\mathbf{x}}^n(i,j)$	$K \times 1$	k	$\hat{\mathbf{x}}^n(i,j) = \mathbf{x}^n(i,j) - \bar{\mathbf{x}}$ with zero mean
$Cov(\hat{\mathbf{x}})$	$K \times K$	-	covariance matrix for $\{\hat{\mathbf{x}}^n(i,j)\}$
ξ^*	$K \times 1$	k	first eigenvector of $Cov(\hat{\mathbf{x}})$

layer can be easily obtained using a standard classification network (*e.g.* VGGNet [69]). See Table 3 for more details.

Utilizing images with only keywords labeling, the CAM technique aims at producing a class specific attention map $\mathbf{M}_c^n$ for each class c using the **feature maps** $\{\mathbf{X}_k^n\}$:

$$\mathbf{M}_c^n = \sum_{k=1}^{K} \omega_k^c \mathbf{X}_k^n, \tag{1}$$

where the weights ω^c could be trained using keyword level weak supervision [119]. Notice that each spatial element of the class activation map $\mathbf{M}_c^n$ can be independently estimated using the weights ω^c and the channel-wise **descriptor** in $\mathbf{X}^n$ at spatial location (i,j) as

$$\mathbf{M}_c^n(i,j) = (\omega^c)^\top \cdot \mathbf{x}^n(i,j). \tag{2}$$

Thus the CAM [119] technique essentially plays a linear transformation that transforms the image features $\mathbf{x}^n(i,j)$ into class specific activation scores $\mathbf{M}_c^n(i,j)$ using the learned class specific weights ω^c.

Unfortunately, in the co-saliency detection problem settings, the keywords level supervision is not available. Thus, we have to discover the weighting ω for the common objects in an unsupervised fashion, by revealing the internal structure of the image features. Ideally, the unknown common object category among a group of associated images $\{\mathbf{I}^n\}_{n=1}^N$ should corresponds to a linear projection that results in high class activation scores in the common object regions, while having low class activation scores in other image regions. From another point of view, the common object category should correspond to the linear transformation that generates the highest variance (most informative) in the resulting class activation maps. Follow the idea in coarse localization task [123], we achieve this gold by exploring the classical principle component analysis (PCA) [124], which is the simplest way of revealing the internal structure of the data in a way that best explains the variance in the data.

Specifically, given the associated images $\{\mathbf{I}^n\}_{n=1}^N$, with corresponding feature activations $\mathbf{X}^n$ for each image $\mathbf{I}^n$, we aims at finding the linear transformation of $\mathbf{X}^n$ that results in the co-attention maps $\{\mathbf{A}^n\}$ with the highest variance. This can be achieved by analyzing the covariance matrix of the feature descriptors $\{\mathbf{x}^n(i,j)\}$. Let $\bar{\mathbf{x}} = \frac{1}{Z}\sum_n \sum_{i,j} \mathbf{x}^n(i,j)$, where $Z = N \times H \times W$. We have the zero mean version of the descriptors as $\hat{\mathbf{x}}^n(i,j) = \mathbf{x}^n(i,j) - \bar{\mathbf{x}}$. The covariance matrix can be denoted as

$$Cov(\hat{\mathbf{x}}) = \frac{1}{Z}\sum_n \sum_{i,j} (\hat{\mathbf{x}}^n(i,j) - \bar{\mathbf{x}})(\hat{\mathbf{x}}^n(i,j) - \bar{\mathbf{x}})^T. \tag{3}$$

Then the expected linear projection can be established by using the eigenvector ξ^*, that corresponds to the largest eigenvalue of $Cov(\hat{\mathbf{x}})$. Thus, the co-attention projection can be designed as a projection that presents the features in its most informative viewpoint

$$\mathbf{A}^n(i,j) = \xi^{*\top} \cdot \hat{\mathbf{x}}^n(i,j). \tag{4}$$

TABLE 4
Benchmarking results of 18 leading CoSOD approaches on two classical [40], [41], and our CoSOD3k. The symbol "○" means that the code or results are not available. Note that the UMLF adopts half of the images from both MSRC and CoSal2015 to train their model. Underline indicates the scores generated by models (*e.g.*, SP-MIL and UMLF) that have been trained on corresponding dataset. See Table 2 for more training details.

	Metric	CBCS [29]	ESMG [53]	RFPR [120]	CSHS [51]	SACS [57]	CODR [121]	UMLF [72]	DIM [58]‡	CODW [59]‡	MIL [64]‡	IML [85]‡	GONet [75]‡	SP-MIL [63]‡	CSMG [95]‡	CPD [122]‡	GSPA [101]‡	AGC [103]‡	EGNet [118]‡	***CoEG-Net*** Ours‡
iCoSeg	$E_\phi \uparrow$	.797	.784	.841	.841	.817	.889	.827	.864	.832	.799	.895	.864	.843	.889	.900	.818	.897	.911	**.912**
	$S_\alpha \uparrow$	.658	.728	.744	.750	.752	.815	.703	.758	.750	.727	.832	.820	.771	.821	.861	.784	.821	.875	**.875**
	$F_\beta \uparrow$	.705	.685	.771	.765	.770	.823	.761	.797	.782	.741	.846	.832	.794	.850	.855	.718	.837	.875	**.876**
	$\epsilon \downarrow$	.172	.157	.170	.179	.154	.114	.226	.179	.184	.186	.104	.122	.174	.106	**.057**	.098	.079	.060	.060
CoSal2015	$E_\phi \uparrow$	.656	.640	○	.685	.749	.749	.769	.695	.752	.720	-	.805	○	.842	.841	.855	**.890**	.843	.882
	$S_\alpha \uparrow$	.544	.552	○	.592	.694	.689	.662	.592	.648	.673	-	.751	○	.774	.814	.797	.823	.818	**.836**
	$F_\beta \uparrow$	.532	.476	○	.564	.650	.634	.690	.580	.667	.620	-	.740	○	.784	.782	.779	.831	.786	**.832**
	$\epsilon \downarrow$	.233	.247	○	.313	.194	.204	.271	.312	.274	.210	-	.160	○	.130	.098	.099	.090	.099	**.077**
CoSOD3k	$E_\phi \uparrow$	.637	.635	○	.656	○	.700	.758	.662	○	○	.773	○	○	.804	.791	.800	.823	.793	**.825**
	$S_\alpha \uparrow$	.528	.532	○	.563	○	.630	.632	.559	○	○	.720	○	○	.711	.757	.736	.759	.762	**.762**
	$F_\beta \uparrow$	.466	.418	○	.484	○	.530	.639	.495	○	○	.652	○	○	.709	.699	.682	.729	.702	**.736**
	$\epsilon \downarrow$	.228	.239	○	.309	○	.229	.285	.327	○	○	.164	○	○	.157	.120	.124	.094	.119	**.092**

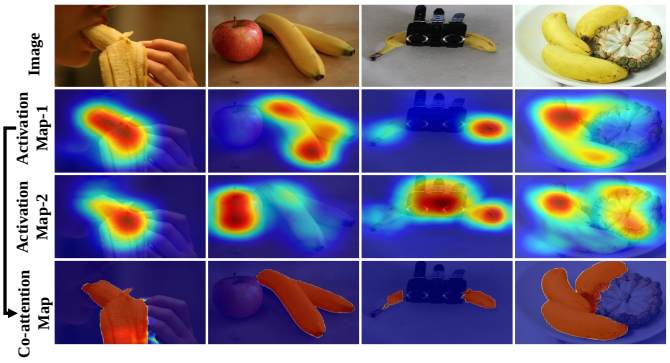


Fig. 10. Visualization of the common activation maps (second and third row), using largest and second eigenvalue, and their corresponding post-processed (*i.e.*, manifold ranking and DenseCRF) co-attention map $\mathbf{A}^n$ (fourth row) selected from the "banana" group of CoSal2015 [40].

Fig. 10 shows some visual templates of common activation maps (second and third row) resulting from Eq. 4. The given images contains multiple objects of diverse categories including banana, apple, bottle and pineapple, increasing the difficulty of differentiate correct regions, while using the largest eigenvalue of $Cov(\hat{\mathbf{x}})$ (second row) can sufficiently localize the common objects and mask them out (last row) initially.

4.3 Implementation

The VGGNet16 network [69] after removing the top layer is selected as our backbone for a fair comparison. The training process is finished in 30 epochs and the learning rate is divided by 10 after 15 epochs. For the edge-guided contextual saliency network, the setting is the same with [118]. Note that in the training stage, the loss function is the same with EGNet. Similar to the post-processing in [75], we utilize the DenseCRF [47] and manifold ranking [125] to further refine co-attention map $\mathbf{A}^n$ before integrating them with the saliency prior map $\mathbf{S}^n$. The examples are shown in the third row of Fig. 10.

5 Benchmark Experiments

5.1 Experimental Settings

• **Evaluation Metrics.** To provide a comprehensive evaluation, four widely used metrics are employed for evaluating CoSOD performance, including maximum F-measure F_β [37], mean absolute error (MAE) ϵ [36], S-measure S_α [126], and maximum E-measure E_ϕ [127]. The complete evaluation toolbox can be found at https://github.com/DengPingFan/CoSODToolbox.

F-measure F_β [37] evaluate the weighted harmonic mean of precision and recall. The saliency maps have to be binarized using different threshold, where each threshold corresponds to a binary saliency prediction. The predicted and ground-truth binary maps are compared to get precision and recall values. F_β is typically chosen as the F-measure score that corresponds to the best fixed threshold for the whole dataset.

MAE ϵ [36] is a much simple evaluation metric that directly measures the absolute difference between the ground-truth value and the predicted value, without any binarization requirements. Both F-measure and MAE evaluate the prediction in a pixel by pixel manner.

S-measure S_α [126] is designed to evaluate the structural similarity between a saliency map and the corresponding ground-truth. It can directly evaluate the continuous saliency prediction without binarization and consider the large scale structure similarity at the same time.

E-measure E_ϕ [127] is a perceptual metric that evaluates both local and global similarity between the predicted map and ground-truth simultaneously.

• **Competitors.** In the CoSOD experiments, we evaluate/compare sixteen SOTA CoSOD models, including seven traditional methods [29], [51], [53], [57], [72], [120], [121] and nine deep learning models [58], [59], [63], [64], [75], [85], [95], [118], [122]. The methods were chosen based on two criteria: (1) representative, and (2) released code or results.

• **Benchmark Protocols.** We evaluate on two existing

Fig. 11. Examples of our CoSOD3k. We visualize segmentation examples for representative object categories from 13 super-classes.

TABLE 5
Per super-class average E-measure performance E_ϕ on our CoSOD3k. Vege. = Vegetables, Nece. = Necessary, Traf. = Traffic, Cosm.= Cosmetic, Inst. = Instrument, Kitch. = Kitchenware, Elec. = Electronic, Anim. = Animal, Oth. = Others. "All" means the score on the whole dataset. We only evaluate the 10 state-of-the-art models with released codes. Note that CPD and EGNet are the top-2 SOD models on the socbenchmark (http://dpfan.net/socbenchmark).

	Vege.	Food	Fruit	Tool	Nece.	Traf.	Cosm.	Ball	Inst.	Kitch.	Elec.	Anim.	Oth.	All
#Sub-class	4	5	9	11	12	10	4	7	14	9	9	49	17	160
ESMG [53]	.577	.635	.735	.625	.546	.673	.633	.559	.655	.631	.629	.687	.592	.635
CBCS [29]	.680	.621	.739	.617	.603	.666	.664	.619	.627	.625	.640	.672	.594	.637
CSHS [51]	.613	.591	.733	.677	.585	.691	.677	.563	.637	.651	.665	.715	.624	.656
CODR [121]	.682	.682	.774	.679	.634	.756	.678	.580	.671	.686	.695	.771	.638	.700
DIM‡ [58]	.622	.687	.773	.650	.604	.708	.633	.577	.665	.612	.641	.709	.623	.662
UMLF [72]	.781	.777	.781	.694	.779	.836	.714	.668	.711	.763	.748	.810	.690	.758
IML‡ [85]	.802	.725	.808	.740	.714	.867	.753	.653	.734	.795	.729	.855	.663	.773
CPD‡ [122]	.805	.763	.818	.734	.758	.894	.763	.629	.638	.848	.784	.892	.693	.791
EGNet‡ [118]	**.833**	.761	.815	.746	.767	.890	.769	.632	.654	.841	.771	**.893**	.697	.793
CSMG‡ [95]	.755	**.872**	**.854**	.722	.744	**.908**	.766	.778	.690	**.849**	.840	.885	.690	.804
CoEG-Net (Ours)‡	.802	.842	.840	**.811**	**.790**	.897	**.795**	**.780**	**.746**	.844	**.842**	.881	**.739**	**.825**

CoSOD datasets, *i.e.*, *iCoSeg* [41], and *CoSal2015* [40], and our CoSOD3k. To the best of our knowledge, ours is the largest-scale and most comprehensive benchmark. For comparison, we run the available codes directly, either under default settings (*e.g.*, CBCS [29], ESMG [53], RFPR [120], CSHS [51], SACS [57], CODR [121], UMLF [72], DIM [58], CPD [122], and EGNet [118]) or using the CoSOD maps provided by the authors (*e.g.*, IML [85], CODW [59], GONet [75], SP-MIL [63], and CSMG [95]).

5.2 Quantitative Comparisons

5.2.1 Performance on iCoSeg.

The iCoSeg dataset [41] was originally designed for image co-segmentation but is widely used for the CoSOD task. Interestingly, as can be seen in Table 4, the two SOD models (*i.e.*, EGNet [118] and CPD [122]) achieve the state-of-the-art performances. The CoSOD methods (*e.g.*, CODR [121], IML [85], and CSMG [95]) also obtain very close performances to the top SOD models (*i.e.*, EGNet [118] and CPD [122]). Our *CoEG-Net* obtains the best performance in E_ϕ, S_α, and F_β, but the results are very close to those of the backbone, *i.e.*, EGNet [118]. One possible reason is that the iCoSeg dataset contains a lot of images with single objects, which can easily be detected by SOD models. The co-salient feature is not an importance role in iCoSeg dataset. This also suggests that the iCoSeg dataset may not be suitable for evaluating CoSOD methods in the deep learning era. Some examples can be found in Fig. 12.

5.2.2 Performance on CoSal2015.

Table 4 shows the evaluation results on the CoSal2015 dataset [40]. One interesting observation is that the existing salient object detection methods, *e.g.*, EGNet [118] and CPD [122], obtain higher performances than most CoSOD methods. This implies that some top-performing salient object detection frameworks may be better-suited for extension to CoSOD tasks. The CoSOD method CSMG [95] achieves comparable performance in E_ϕ (0.842) and F_β (0.784), but worse scores in S_α (0.774) and ϵ (0.130). This demonstrates that existing CoSOD methods cannot solve the task well. Our *CoEG-Net* obtains the best results, significantly outperforming both SOD and Co-SOD baselines.

5.2.3 Performance on CoSOD3k.

The overall results on our CoSOD3k are presented in Table 4. As expect, our model still achieve the best performance. To provide deeper insight into each group, we report the performances of models on 13 super-classes in Table 5. We observe that lower average scores are achieved on classes such as Other (*e.g.*, *baby bed* and *pencil box*), Instrument (*e.g.*, *piano*, *guitar*, *cello*, *etc.*), Necessary (*e.g.*, *pitcher*), Tool (*e.g.*, *axe*, *nail*, *chain saw*, *etc.*), and Ball (*e.g.*, *soccer*, *tennis*, *etc.*), which contain complex structures in real scenes. Note that almost all of the deep-based models (*e.g.*, EGNet [118], CPD [122], IML [85], and CSMG [95]) perform better than the traditional approaches (CODR [121], CSHS [51], CBCS [29], and ESMG [53]), demonstrating the potential advantages in utilizing deep learning techniques to address the CoSOD problem. Another interesting finding is that edge features can help provide good boundaries for the results. For instance, the best methods from both traditional (CSHS [51]) and deep learning models (*e.g.*, EGNet [118]) introduce edge information to aid detection. Finally, our method *CoEG-Net* obtains the best performance on average, with an E_ϕ of 0.825 which is much higher than the second-best method, *i.e.*, CSMG [95] with 0.804. Moreover, the performances (Table 4) of all methods are worse than on the other two datasets (*e.g.*, iCoSeg and CoSal2015), which clearly shows

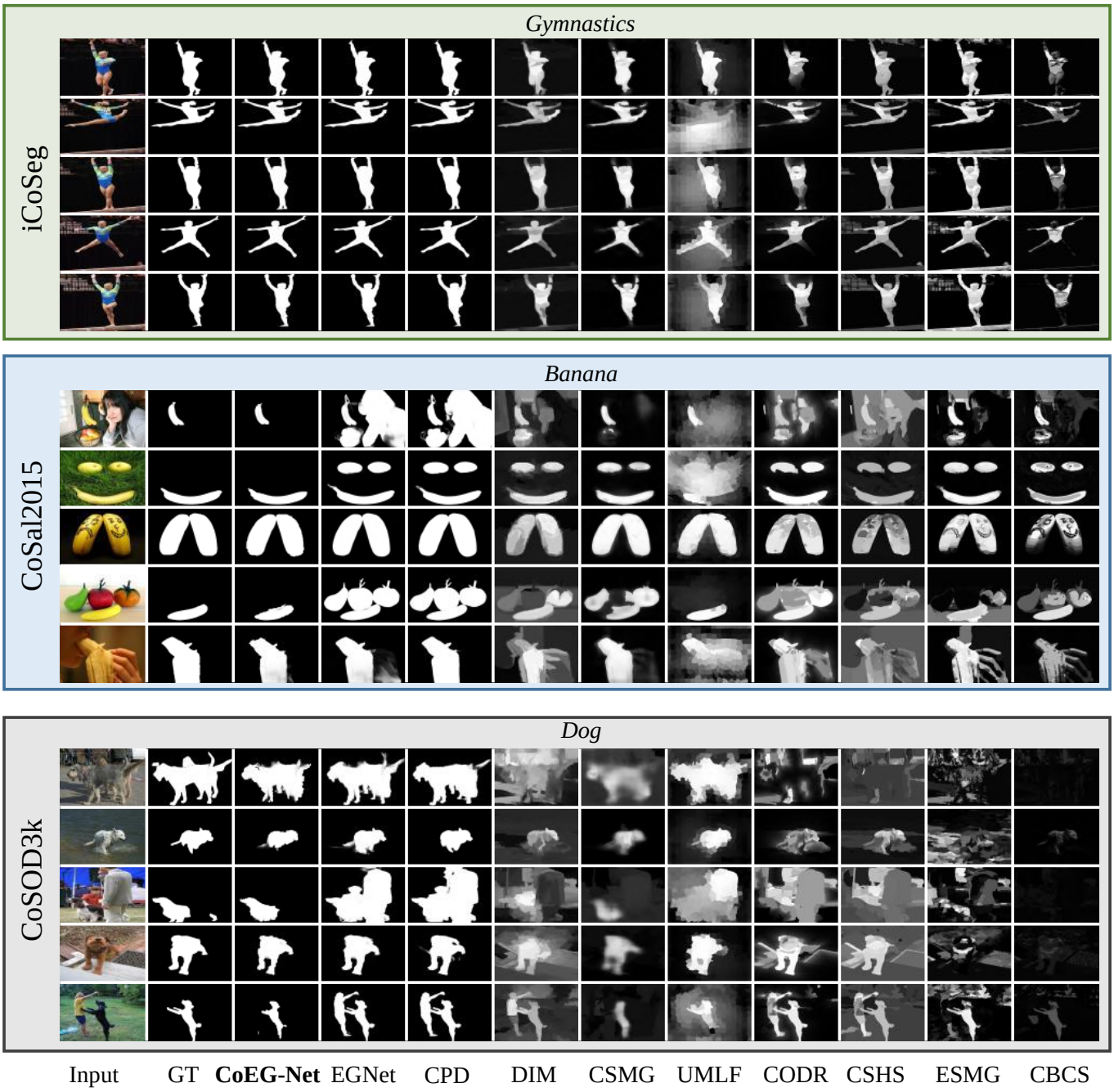


Fig. 12. Qualitative examples of 10 representative models evaluated on iCoSeg [41], CoSal2015 [40], and our CoSOD3k.

that the proposed CoSOD3k dataset is challenging and leaves abundant room for further research.

5.3 Qualitative Comparisons

Fig. 12 shows some qualitative examples on iCoSeg, CoSal2015, and our CoSOD3k. As can be seen, the SOD models, *e.g.*, EGNet [118] and CPD [122], detect all salient objects and obtain sharp boundaries, performing better than other baselines. However, these SOD models ignore the context information.

For example, the "banana" group in the CoSal2015 dataset contain several other irrelevant objects, *e.g.*, oranges, pineapples, and apples. The SOD models cannot distinguish these as being irrelevant. Another similar situation also occurs in the images in the dog group of our CoSOD3k, where the humans (the third and fifth images) are detected together with the dogs. On the other hand, the CoSOD methods, *e.g.*, CSMG [95] and DIM [58], can identify the common salient objects and remove the other objects (*e.g.*, human). However, these CoSOD methods cannot produce accurate predicted maps, especially around object boundaries. By contrast, our *CoEG-Net* preserves the advantages of SOD and CoSOD methods, and obtains the best visual results in all datasets.

5.4 Comparison with Baselines

Our baseline *CoEG-Net* consists of a co-attention projection and a basic SOD model. In order to explore the efficiency

TABLE 6
Ablative studies of our model on three benchmark datasets, where Ours-A, Ours-P, Ours-E represent the co-salient results of Amulet, PiCANet, EGNet on our baseline, respectively.

Datasets	Metric	Amulet	**Ours-A**	PiCANet	**Ours-P**	EGNet	**Ours-E**
iCoSeg	$E_\phi \uparrow$	.877	**.878**	.906	**.907**	.911	**.912**
	$S_\alpha \uparrow$	.828	**.829**	.869	**.870**	.875	**.875**
	$F_\beta \uparrow$	.829	**.829**	.854	**.854**	.875	**.876**
	$\epsilon \downarrow$	.088	**.087**	.065	**.064**	.060	**.060**
CoSal2015	$E_\phi \uparrow$	.772	**.831**	.859	**.870**	.843	**.882**
	$S_\alpha \uparrow$	.719	**.744**	.801	**.825**	.818	**.836**
	$F_\beta \uparrow$	.684	**.758**	.799	**.818**	.786	**.832**
	$\epsilon \downarrow$	.147	**.125**	.090	**.084**	.099	**.077**
CoSOD3k	$E_\phi \uparrow$	.752	**.803**	.780	**.819**	.793	**.825**
	$S_\alpha \uparrow$	.685	**.692**	.750	**.758**	.762	**.762**
	$F_\beta \uparrow$	.629	**.700**	.682	**.724**	.702	**.736**
	$\epsilon \downarrow$	.145	**.122**	.137	**.095**	.119	**.092**

TABLE 7
Average running time of ten SOTA models.

Models	CBCS [29]	ESMG [53]	CSHS [51]	CODR [121]
Time (seconds)	0.3	1.2	102	35
Language	Matlab	Matlab	Matlab	Matlab
Models	UMLF [72]	DIM‡ [58]	CSMG‡ [95]	CPD‡ [122]
Time (seconds)	87	25	3.2	0.016
Language	Matlab	Matlab	Caffe	PyTorch
Models	EGNet‡ [118]	Ours‡		
Time (seconds)	0.034	2.3		
Language	PyTorch	PyTorch		

of the co-attention projection, we (1) adopt the same training dataset (*i.e.*, DUTS [30]) and test datasets (*i.e.*, iCoSeg, CoSal2015, and CoSOD3k) for three SOTA SOD models (*i.e.*, Amulet [17], PiCANet [92], and EGNet [118]); and (2) apply the same co-attention projection strategy for these models, as presented in Section 4, to conduct this experiment. Table 6 shows the performances of three baselines in terms of E_ϕ, S_α, F_β, and ϵ metrics. Based on the results, we observe that: (i) On the relatively simple iCoSeg dataset, our baselines (*i.e.*, Ours-A/-P/-E) slightly improve upon the backbone models (*i.e.*, Amulet, PiCANet, and EGNet). We note that because this dataset contains a large number of single objects with similar appearances (Fig. 12) in each group, only using a SOD model can achieve very high performance. This conclusion is consistent with the analysis in Section 5.2.1; (ii) On the classical CoSal2015 dataset, our baselines are consistently better than the backbones in terms of all four metrics. It is worth noting that, for this more complex dataset, we still obtain a 2.5%, 1.4%, and 1.8% S_α score improvement; (iii) For the proposed and most challenging dataset CoSOD3k, we find that the improvement is still significant (*e.g.*, 7.1% F_β score for Amulet). To further analyze the improvement, we also provide the 160 sub-class performances in the supplementary materials. We observe that, for objects in the common super-class (*i.e.*, 'Ball') such as "rugby_ball" and "soccer_ball", we achieve 23.5% and 23.9% F_β improvements. We attribute this to the co-attention projection operation being able to automatically learn mutual-features, which are crucial for overcoming challenging ambiguities.

5.5 Running Time

Our *CoEG-Net* is implemented in PyTorch and Caffe with an RTX 2080Ti GPU for acceleration. For traditional algorithms (CBCS [29], ESMG [53], CSHS [51], CODR [121], and UMLF [72]), the comparison experiments are executed on a laptop with Inter(R) Core(TM) i7-2600 CPU @3.4GHz. The remaining deep learning models (DIM [58], CSMG [95], CPD [122], and EGNet [118]) are tested on a workstation with Intel(R) Core(TM) i7-8700K CPU @3.70GHz and an RTX 2080Ti GPU. As shown in Table 7, among the top-3 CoSOD models, *i.e.*, the proposed *CoEG-Net*, CSMG [95], and UMLF [72], evaluated in terms of E_ϕ measure on the proposed CoSOD3k, our model achieves the fastest inference time. In addition, compared with the top-2 fastest CoSOD models (*i.e.*, CBCS [29] and ESMG [95]), although the proposed model has a longer test time, it obtains a significantly improved S_α measure. This partially suggests that our framework is not only efficient but also effective for the CoSOD task. However, compared to two recently released state-of-the-art models, CPD [122] and EGNet [118], there is still large room for improvement in running time.

6 Discussion and Future Directions

From the evaluation, we observe that, in most cases, the current SOD methods (*e.g.*, EGNet [118] and CPD [122]) can obtain very competitive or even better performances than the CoSOD methods (*e.g.*, CSMG [95] and SP-MIL [63]). However, this does not necessarily mean that the current datasets are not complex enough or using the SOD methods directly can obtain the good performances—the performances of the SOD methods on the CoSOD datasets are actually lower than those on the SOD datasets. For example, EGNet achieves 0.937 and 0.943 F_β scores on the HKU-IS dataset [113] and ECSSD dataset [116], respectively. However, it only obtains 0.786 and 0.702 F_β scores on the CoSal2015 and CoSOD3k datasets, respectively. Consequently, the evaluation results reveal that many problems in CoSOD are still under-studied and this makes the existing CoSOD models less effective. In this section, we discuss four important issues (*i.e.*, scalability, stability, compatibility, and metrics) that have not been fully addressed by the existing co-salient object detection methods and should be studied in the future. Finally, we discuss the weakness of the the proposed *CoEG-Net* framework.

• **Scalability.** The scalability is one of the most important issues that needs to be considered when designing CoSOD algorithms. Specifically, it indicates the capability of a CoSOD model of handling large-scale image scenes. As we know, one key property of CoSOD is that the model needs to consider multiple images from each group. However, in reality, an image group may contain numerous related images. Under this circumstance, methods that do not consider scalability would have huge computational costs and take a very long time to run, making them unacceptable in practice (*e.g.*, CSHS-102s and UMLF-87s). Thus, how to address the scalability issue, or how to reduce the computational complexity caused by the number of images contained in an image group, becomes a key problem in this field, especially when applying CoSOD methods for real-world applications.

• **Stability.** Another important issue is the stability of model. When dealing with image groups containing multiple images, some existing methods (*e.g.*, HCNco [128], PCSD [48], and IPCS [32]) divide the image group into image pairs or image sub-groups (*e.g.*, GD [67]). Another school of methods adopt the RNN-based model (*e.g.*, GWD [99]), which involves assigning an order to the input images. These strategies all make the overall training process unstable as there is no principle way of dividing image groups or assigning input order to related images. In other words, when generating image sub-groups or assigning the input orders following different strategies, the learning procedure produces different co-saliency detectors, and the test results are also unstable. Consequently, this not only brings difficulty for evaluating the performance of the learned co-saliency detectors but also influences the application of the co-salient object detection.

• **Compatibility.** Introducing SOD in CoSOD is a direct yet effective strategy for building CoSOD framework as the single image saliency can conduce to the co-saliency pattern identification. However, most existing CoSOD works only utilize the results or features of the SOD models as useful information cues. The proposed *CoEG-Net* baseline still follows this two-stage framework that spends more inference time than the single SOD model. Although as a preliminary attempt, we have also achieved the best performance among the existing CoSOD models. From this point of view, one further direction for leveraging the SOD technique is to deeply combine a CNN-based SOD network with a CoSOD model to build an end-to-end trainable framework for detecting CoSOD directly. To achieve this goal, one needs to consider the compatibility of the CoSOD framework, making it convenient for integrating the existing SOD techniques.

• **Metrics.** Current evaluation metrics for CoSOD are designed in terms of SOD, *i.e.*, they calculate the mean of the SOD scores on each group directly. In contrast to SOD, CoSOD involves relationship information between co-salient objects of different images, which is more important for CoSOD evaluation. For example, current CoSOD metrics assume taht the target objects have similar sizes in all images. As the objects actually have different sizes in different images, these metrics ($S_\alpha, E_\phi, F_\beta, \epsilon$ in Sec. 5) would likely be inclined to detecting large objects. Moreover, the current CoSOD metrics are based towards detecting objects in a single image, rather than identifying co-occurring objects across multiple images. Thus, how to design suitable metrics for CoSOD is an open issue.

• **Weakness.** Compared with the end-to-end CoSOD detection frameworks that output binary predictions with smoothed fine-structures, the prediction results of *CoEG-Net* suffers from coarse boundaries, indicating that *CoEG-Net* cannot finely preserve detailed shape information for the co-salient objects. Some failed detection cases are shown in Fig. 13.

• **Potential Applications.** In this part, we discuss two potential new applications that could benefit from the high-quality CoSOD models. For more CoSOD applications, please refer to the related survey in [18], [34].

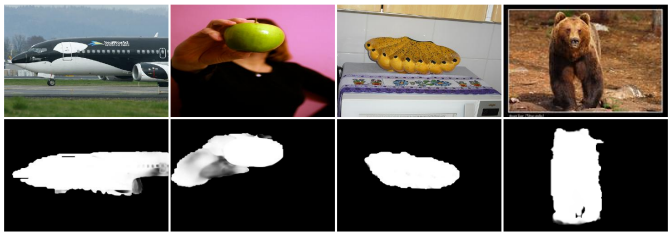


Fig. 13. Some challenge cases for our *CoEG-Net*.

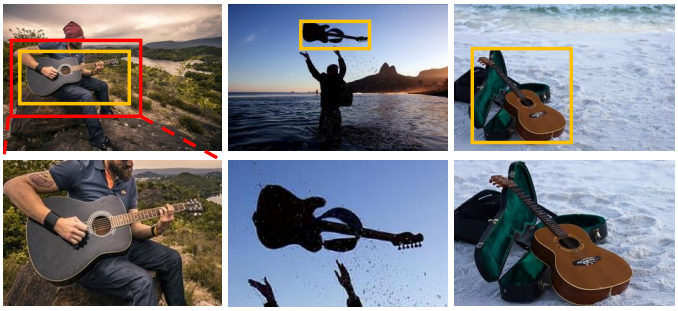


Fig. 14. Collection-Aware Crops.

Collection-Aware Crops. This application is derived from Jacob *et al.* 's work [23]. It studies where people look when comparing images and triggers the seminal works on the CoSOD task. Sharing the same spirit, we show a more general potential application which is not limited to image pairs. As an example, when dealing with the automatic thumb-nailing task as in Fig. 14, we first obtain the yellow bounding box from the saliency maps generated by our *CoEG-Net*. After that, an enlarged (∼60 pixels) red box[3] is used to identify the crop regions automatically. To obtain high-quality crops from the first row, we can also introduce existing SOTA super-resolution techniques [129], [130] to further improve the visualization results.

Object Co-Localization. As shown by DeepCO[3] [97], the co-saliency detection results will provide the class-agnostic attention cues for the object co-location task. Introducing our *CoEG-Net* to existing commerce application will be a possible solution to improve the performance in this field.

7 Conclusion

In this paper, we have presented a comprehensive investigation on the co-salient object detection (CoSOD) task. After identifying the serious data bias in current dataset, which assume that each image group contains salient object(s) of similar visual appearance, we built a new high-quality dataset, named CoSOD3k, containing co-salient object(s) that are similar at a semantic or conceptual level. Notably, CoSOD3k is the most challenging CoSOD dataset so far, containing 160 groups and total of 3,316 images labeled with category, bounding box, object-level, and instance-level annotations. Our CoSOD3k dataset makes a significant leap in terms of diversity, difficulty and scalability, benefiting several related vision tasks, *e.g.*, co-segmentation, weakly

3. Note that we keep the original width of the yellow box when the enlarged red box touched the boundary of the image.

supervised localization, and instance-level detection, and their future development.

To creat an effective co-salient object detector, we integrated existing SOD techniques to build a unified, trainable CoSOD framework called *CoEG-Net*. Specifically, we augmented our prior model EGNet with a co-attention projection strategy to enable efficient common information learning, improving the scalability and stability of the co-salient object detection framework.

Besides, this paper has also provided a comprehensive study by summarizing 40 cutting-edge algorithms, benchmarking 18 of them over two classical datasets, as well as the proposed CoSOD3k. By evaluating recent SOD and CoSOD methods, this paper demonstrated that the SOD methods are surprisingly better. This is an interesting finding that can guide further investigation into better CoSOD algorithms. We hope the studies presented in this work will give a strong boost to the growth of the CoSOD community. In the future, we plan to increase the dataset scale to spark more novel ideas.

ACKNOWLEDGMENT

We also thank professor Kaihua Zhang from Nanjing University of Information Science & Technology for insightful feedback. This research was supported by NSFC (61922046), S&T innovation project from Chinese Ministry of Education, and Tianjin Natural Science Foundation (18ZXZNGX00110).

REFERENCES

[1] D.-P. Fan, Z. Lin, G.-P. Ji, D. Zhang, H. Fu, and M.-M. Cheng, "Taking a Deeper Look at the Co-salient Object Detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[2] 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 *Eur. Conf. Comput. Vis.*, 2018, pp. 186–202.

[3] Y. Zeng, P. Zhang, J. Zhang, Z. Lin, and H. Lu, "Towards high-resolution salient object detection," in *Int. Conf. Comput. Vis.*, 2019, pp. 7234–7243.

[4] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, and S. Kwong, "Nested network with two-stream pyramid for salient object detection in optical remote sensing images," *IEEE T. Geosci. Remote. Sens. Lett.*, vol. 57, no. 11, pp. 9156–9166, 2019.

[5] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Comput. Vis. Media*, vol. 5, no. 2, pp. 117–150, 2019.

[6] X. Qin, D.-P. Fan, C. Huang, C. Diagne, Z. Zhang, A. C. Sant'Anna, A. Suàrez, M. Jagersand, and L. Shao, "Boundary-aware segmentation network for mobile and web applications," *arXiv preprint arXiv:2101.04704*, 2021.

[7] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for rgbd salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3927–3936.

[8] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking RGB-D Salient Object Detection: Models, Datasets, and Large-Scale Benchmarks," *IEEE T. Neural Netw. Learn. Syst.*, 2020.

[9] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. Saleh, S. Aliakbarian, and N. Barnes, "Uncertainty inspired rgb-d saliency detection," *arXiv preprint arXiv:2009.03075*, 2020.

[10] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for rgb-d salient object detection and beyond," *arXiv preprint arXiv:2008.12134*, 2020.

[11] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "Rgb-d salient object detection: A survey," *Comput. Vis. Media*, pp. 1–33, 2021.

[12] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 8554–8564.

[13] R. Cong, J. Lei, H. Fu, F. Porikli, Q. Huang, and C. Hou, "Video Saliency Detection via Sparsity-Based Reconstruction and Propagation," *IEEE T. Image Process.*, vol. 28, no. 10, pp. 4819–4831, 2019.

[14] W. Wang, J. Shen, J. Xie, M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE T. Pattern Anal. Mach. Intell.*, 2020.

[15] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2007.

[16] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 478–487.

[17] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Int. Conf. Comput. Vis.*, 2017, pp. 202–211.

[18] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE T. Circuit Syst. Video Technol.*, vol. 29, no. 10, pp. 2941–2959, 2018.

[19] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *arXiv preprint arXiv:1904.09146*, 2019.

[20] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," in *Int. Conf. Comput. Vis.*, 2019, pp. 3799–3808.

[21] H. Bi, K. Wang, D. Lu, C. Wu, W. Wang, and L. Yang, "C2net: a complementary co-saliency detection network," *The Vis. Comput.*, pp. 1–13, 2020.

[22] J. Ren, Z. Liu, G. Li, X. Zhou, C. Bai, and G. Sun, "Co-saliency detection using collaborative feature extraction and high-to-low feature integration," in *Int. Conf. Multimedia and Expo*, 2020, pp. 1–6.

[23] D. E. Jacobs, D. B. Goldman, and E. Shechtman, "Cosaliency: Where people look when comparing images," in *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 2010, pp. 219–228.

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

[25] H. Fu, D. Xu, B. Zhang, S. Lin, and R. K. Ward, "Object-Based Multiple Foreground Video Co-Segmentation via Multi-State Selection Graph," *IEEE T. Image Process.*, vol. 24, no. 11, pp. 3415–3424, 2015.

[26] L. Zhang, J. Zhang, Z. Lin, H. Lu, and Y. He, "Capsal: Leveraging captioning to boost semantics for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 6024–6033.

[27] G. Liu and D. Fan, "A model of visual attention for natural image retrieval," in *IEEE Int. Conf. Inf. Sci. Cloud Comput. Companion*, 2013, pp. 728–733.

[28] M.-M. Cheng, N. J. Mitra, X. Huang, and S.-M. Hu, "Salientshape: group saliency in image collections," *Vis. Comput.*, vol. 30, no. 4, pp. 443–453, 2014.

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

[30] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 136–145.

[31] H. Yu, K. Zheng, J. Fang, H. Guo, W. Feng, and S. Wang, "Co-saliency detection within a single image," in *AAAI Conf. Art. Intell.*, 2018, pp. 7509–7516.

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

[33] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Int. Conf. Comput. Vis.*, 2005, pp. 1800–1807.

[34] D. Zhang, H. Fu, J. Han, A. Borji, and X. Li, "A review of co-saliency detection algorithms: Fundamentals, applications, and challenges," *ACM Trans Intell Syst Technol*, vol. 9, no. 4, pp. 1–31, 2018.

[35] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 1521–1528.

[36] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *Int. Conf. Comput. Vis.*, 2013, pp. 1529–1536.

[37] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 1597–1604.

[38] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 248–255.

[39] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE T. Image Process.*, vol. 24, no. 12, pp. 5706–5722, 2015.

[40] D. Zhang, J. Han, C. Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 2994–3002.

[41] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "icoseg: Interactive co-segmentation with intelligent scribble guidance," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 3169–3176.

[42] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[43] J. Dai, Y. Nian Wu, J. Zhou, and S.-C. Zhu, "Cosegmentation and cosketch by unsupervised learning," in *Int. Conf. Comput. Vis.*, 2013, pp. 1305–1312.

[44] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3166–3173.

[45] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global Contrast based Salient Region Detection," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, 2015.

[46] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Eur. Conf. Comput. Vis.*, 2016, pp. 825–841.

[47] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Adv. Neural Inform. Process. Syst.*, 2011, pp. 109–117.

[48] H.-T. Chen, "Preattentive co-saliency detection," in *IEEE Int. Conf. Image Process.*, 2010, pp. 1117–1120.

[49] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Adv. Neural Inform. Process. Syst.*, 2009, pp. 681–688.

[50] H. Li, F. Meng, and K. N. Ngan, "Co-salient object detection from multiple images," *IEEE T. Multimedia*, vol. 15, no. 8, pp. 1896–1909, 2013.

[51] Z. Liu, W. Zou, L. Li, L. Shen, and O. Le Meur, "Co-saliency detection based on hierarchical segmentation," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 88–92, 2013.

[52] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, 2010.

[53] Y. Li, K. Fu, Z. Liu, and J. Yang, "Efficient saliency-model-guided visual co-saliency detection," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 588–592, 2014.

[54] B. Xu, J. Bu, C. Chen, D. Cai, X. He, W. Liu, and J. Luo, "Efficient manifold ranking for image retrieval," in *ACM Spec. Interest Group Inf. Ret.*, 2011, pp. 525–534.

[55] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE TSMC*, vol. 9, no. 1, pp. 62–66, 1979.

[56] X. Cao, Y. Cheng, Z. Tao, and H. Fu, "Co-saliency detection via base reconstruction," in *ACM Int. Conf. Multimedia*, 2014, pp. 997–1000.

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

[58] D. Zhang, J. Han, J. Han, and L. Shao, "Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining," *IEEE T. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1163–1176, 2015.

[59] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 215–232, 2016.

[60] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 248–255.

[61] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Int. Conf. Learn. Represent.*, 2014.

[62] Y. Bengio *et al.*, "Learning deep architectures for ai," *FTML*, vol. 2, no. 1, pp. 1–127, 2009.

[63] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, 2016.

[64] 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 *Int. Conf. Comput. Vis.*, 2015, pp. 594–602.

[65] P. Siva, C. Russell, T. Xiang, and L. Agapito, "Looking beyond the image: Unsupervised learning for object saliency and detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3238–3245.

[66] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Brit. Mach. Vis. Conf.*, 2014.

[67] L. Wei, S. Zhao, O. E. F. Bourahla, X. Li, and F. Wu, "Group-wise deep co-saliency detection," in *Int. Jt. Conf. Artif. Intell.*, 2017, pp. 3041–3047.

[68] 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 *Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[69] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent.*, may 2015.

[70] X. Yao, J. Han, D. Zhang, and F. Nie, "Revisiting co-saliency detection: A novel approach based on two-stage multi-view spectral rotation co-clustering," *IEEE T. Image Process.*, vol. 26, no. 7, pp. 3196–3209, 2017.

[71] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[72] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based framework for co-saliency detection," *IEEE T. Circuit Syst. Video Technol.*, vol. 28, no. 10, pp. 2473–2483, 2017.

[73] M. Li, S. Dong, K. Zhang, Z. Gao, X. Wu, H. Zhang, G. Yang, and S. Li, "Deep learning intra-image and inter-images features for co-saliency detection," in *Brit. Mach. Vis. Conf.*, 2018, p. 291.

[74] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.

[75] K.-J. Hsu, C.-C. Tsai, Y.-Y. Lin, X. Qian, and Y.-Y. Chuang, "Unsupervised CNN-based co-saliency detection with graphical optimization," in *Eur. Conf. Comput. Vis.*, 2018, pp. 485–501.

[76] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.

[77] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Co-attention cnns for unsupervised object co-segmentation," in *Int. Jt. Conf. Artif. Intell.*, 2018, pp. 748–756.

[78] X. Zheng, Z.-J. Zha, and L. Zhuang, "A feature-adaptive semi-supervised framework for co-saliency detection," in *ACM Int. Conf. Multimedia*, 2018, pp. 959–966.

[79] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 678–686.

[80] C.-C. Tsai, W. Li, K.-J. Hsu, X. Qian, and Y.-Y. Lin, "Image co-saliency detection and co-segmentation via progressive joint optimization," *IEEE T. Image Process.*, vol. 28, no. 1, pp. 56–71, 2018.

[81] D.-j. Jeong, I. Hwang, and N. I. Cho, "Co-salient object detection based on deep saliency networks and seed propagation over an integrated graph," *IEEE T. Image Process.*, vol. 27, no. 12, pp. 5866–5879, 2018.

[82] K. R. Jerripothula, J. Cai, and J. Yuan, "Quality-guided fusion-based co-saliency estimation for image co-segmentation and colocalization," *IEEE T. Multimedia*, vol. 20, no. 9, pp. 2466–2477, 2018.

[83] S. Song, H. Yu, Z. Miao, D. Guo, W. Ke, C. Ma, and S. Wang, "An easy-to-hard learning strategy for within-image co-saliency detection," *Neurocomputing*, vol. 358, pp. 166–176, 2019.

[84] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015.

[85] J. Ren, Z. Liu, X. Zhou, C. Bai, and G. Sun, "Co-saliency Detection via Integration of Multi-layer Convolutional Features and Inter-image Propagation," *Neurocomputing*, vol. 371, pp. 137–146, 2020.

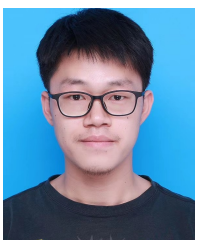
[86] L. Wei, S. Zhao, O. E. F. Bourahla, X. Li, F. Wu, and Y. Zhuang, "Deep group-wise fully convolutional network for co-saliency detection with graph propagation," *IEEE T. Image Process.*, vol. 28, no. 10, pp. 5052–5063, 2019.

[87] B. Li, Z. Sun, L. Tang, Y. Sun, and J. Shi, "Detecting Robust Co-Saliency with Recurrent Co-Attention Neural Network," in *Int. Jt. Conf. Artif. Intell.*, 2019, pp. 818–825.

[88] C. Wang, Z.-J. Zha, D. Liu, and H. Xie, "Robust deep co-saliency detection with group semantic," in *AAAI Conf. Art. Intell.*, 2019, pp. 8917–8924.

[89] B. Jiang, X. Jiang, J. Tang, B. Luo, and S. Huang, "Multiple Graph Convolutional Networks for Co-Saliency Detection," in *Int. Conf. Multimedia and Expo*, 2019, pp. 332–337.
[90] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Int. Conf. Learn. Represent.*, 2017.
[91] 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 *ACM Int. Conf. Multimedia*, 2019, pp. 1375–1382.
[92] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3089–3098.
[93] B. Li, Z. Sun, Q. Wang, and Q. Li, "Co-saliency Detection Based on Hierarchical Consistency," in *ACM Int. Conf. Multimedia*, 2019, pp. 1392–1400.
[94] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Int. Conf. Learn. Represent.*, 2014.
[95] K. Zhang, T. Li, B. Liu, and Q. Liu, "Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3095–3104.
[96] T. Liu, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to Detect A Salient Object," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.
[97] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "DeepCO3: Deep Instance Co-Segmentation by Co-Peak Search and Co-Saliency Detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 8846–8855.
[98] D. Zhang, J. Han, and Y. Zhang, "Supervision by fusion: Towards unsupervised learning of deep salient object detector," in *Int. Conf. Comput. Vis.*, 2017, pp. 4048–4056.
[99] B. Li, Z. Sun, Q. Li, Y. Wu, and A. Hu, "Group-Wise Deep Object Co-Segmentation With Co-Attention Recurrent Neural Network," in *Int. Conf. Comput. Vis.*, 2019, pp. 8519–8528.
[100] G. Gao, W. Zhao, Q. Liu, and Y. Wang, "Co-saliency detection with co-attention fully convolutional network," *IEEE T. Circuit Syst. Video Technol.*, 2020.
[101] Z.-J. Zha, C. Wang, D. Liu, H. Xie, and Y. Zhang, "Robust deep co-saliency detection with group semantic and pyramid attention," *IEEE T. Neural Netw. Learn. Syst.*, 2020.
[102] B. Jiang, X. Jiang, J. Tang, and B. Luo, "Co-saliency detection via a general optimization model and adaptive graph learning," *IEEE T. Multimedia*, 2020.
[103] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 9050–9059.
[104] Z. Zhang, W. Jin, J. Xu, and M.-M. Cheng, "Gradient-induced co-saliency detection," in *Eur. Conf. Comput. Vis.*, 2020.
[105] J. Zhao, R. Bo, Q. Hou, M.-M. Cheng, and P. Rosin, "Flic: Fast linear iterative clustering with active search," *Comput. Vis. Media*, vol. 4, no. 4, pp. 333–348, 2018.
[106] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652–662, 2021.
[107] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[108] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, "PartNet: A Large-scale Benchmark for Fine-grained and Hierarchical Part-level 3D Object Understanding," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 909–918.
[109] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged Object Detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.
[110] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.
[111] H. Jiang, M.-M. Cheng, S.-J. Li, A. Borji, and J. Wang, "Joint Salient Object Detection and Existence Prediction," *Front. Comput. Sci.*, pp. 778—-788, 2017.
[112] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 247–256.
[113] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 5455–5463.
[114] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Int. Conf. Comput. Vis.*, 2001, pp. 416–423.
[115] C. Xia, J. Li, X. Chen, A. Zheng, and Y. Zhang, "What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 4142–4150.
[116] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 1155–1162.
[117] E. L. Kaufman, M. W. Lord, T. W. Reese, and J. Volkmann, "The discrimination of visual number," *Am. J. Psychol.*, vol. 62, no. 4, pp. 498–525, 1949.
[118] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge Guidance Network for Salient Object Detection," in *Int. Conf. Comput. Vis.*, 2019, pp. 8779–8788.
[119] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 2921–2929.
[120] L. Li, Z. Liu, W. Zou, X. Zhang, and O. Le Meur, "Co-saliency detection based on region-level fusion and pixel-level refinement," in *Int. Conf. Multimedia and Expo*, 2014, pp. 1–6.
[121] L. Ye, Z. Liu, J. Li, W.-L. Zhao, and L. Shen, "Co-saliency detection via co-salient object discovery and recovery," *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 2073–2077, 2015.
[122] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3907–3916.
[123] X.-S. Wei, C.-L. Zhang, J. Wu, C. Shen, and Z.-H. Zhou, "Unsupervised object discovery and co-localization by deep descriptor transformation," *Pattern Recognit.*, vol. 88, pp. 113–126, 2019.
[124] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *Philos Mag (Abingdon)*, vol. 2, no. 11, pp. 559–572, 1901.
[125] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Adv. Neural Inform. Process. Syst.*, 2004, pp. 321–328.
[126] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A New Way to Evaluate Foreground Maps," in *Int. Conf. Comput. Vis.*, 2017, pp. 4548–4557.
[127] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment Measure for Binary Foreground Map Evaluation," in *Int. Jt. Conf. Artif. Intell.*, 2018, pp. 698–704.
[128] J. Lou, F. Xu, Q. Xia, W. Yang, and M. Ren, "Hierarchical co-salient object detection via color names," in *IEEE Asian Conf. Pattern Recog.*, 2017, pp. 718–724.
[129] Y. Bahat and T. Michaeli, "Explorable super resolution," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 2716–2725.
[130] A. Lugmayr, M. Danelljan, L. Van Gool, and R. Timofte, "Srflow: Learning the super-resolution space with normalizing flow," in *Eur. Conf. Comput. Vis.*, 2020, pp. 715–732.

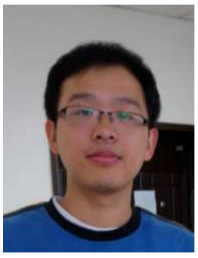
Deng-Ping Fan received his PhD degree from the Nankai University in 2019. He joined Inception Institute of Artificial Intelligence (IIAI) in 2019. He has published about 20 top journal and conference papers such as CVPR, ICCV, etc. His research interests include computer vision, deep learning, and saliency detection, especially on co-salient object detection, RGB salient object detection, RGB-D salient object detection, and video salient object detection.



Tengpeng Li received his BS and MS degree from the School of Information and Control in Nanjing University of Information Science and Technology in Jun. 2017 and Jun. 2020, respectively. He is now a PHD in College of Electronic and Information Engineering, Tongji University. His research interests are computer vision, natural language processing and machine learning.

Zheng Lin is currently a Ph.D. candidate with College of Computer Science, Nankai University, under the supervision of Prof. Ming-Ming Cheng. His research interests include deep learning, computer graphics, machine learning, and computer vision, especially on interactive image segmentation, RGB-D salient object detection, and co-salient object detection.

Ge-Peng Ji is currently a MS of Communication and Information System at School of Computer Science, Wuhan University. His research interests lie in designing deep neural networks and applying deep learning in various fields of low-level vision, such as RGB salient object detection, RGB-D salient object detection, video salient object detection (VSOD), camouflaged object detection (COD), and medical image segmentation.



Dingwen Zhang received the PhD degree from Northwestern Polytechnical University, China, in 2018. He is a professor with the School of Automation, Northwestern Polytechnical University, China. From 2015 to 2017, he was a visiting scholar at the Robotic Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania. His research interests include computer vision and multimedia processing, especially on saliency detection and weakly supervised learning. He is a member of the IEEE.

Ming-Ming Cheng (Senior Member, IEEE) received his PhD degree from Tsinghua University in 2012. He then he did 2 years research fellow, with Prof. Philip Torr in Oxford. He is now a professor at Nankai University, leading the Media Computing Lab. His research interests includes computer graphics, machine learning, computer vision, and image processing. He is an Associate Editor of IEEE TIP. He received several research awards, including the ACM China Rising Star Award, the IBM Global SUR Award, *etc*.

Huazhu Fu is currently a Senior Scientist with Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. He received his Ph.D. from Tianjin University, China, in 2013, and was a Research Fellow with Nanyang Technological University, Singapore for two years. His research interests include computer vision, machine learning, and medical image analysis. He is the Associate Editor of IEEE TMI, IEEE JBHI, *etc*.

Jianbing Shen is currently acting as the Lead Scientist with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates. He is also a Full Professor with the School of Computer Science, Beijing Institute of Technology. He has published about 100 journal and conference papers such as IEEE TPAMI, CVPR, and ICCV. His research interests include computer vision and deep learning. He is an Associate Editor of IEEE TNNLS, IEEE TIP, *etc*.