Taking a Deeper Look at Co-Salient Object Detection

Deng-Ping Fan[1,2,*] Zheng Lin[1,*] Ge-Peng Ji[3] Dingwen Zhang[4] Huazhu Fu[2] Ming-Ming Cheng[1] ✉
[1] TKLNDST, College of CS, Nankai University [2] Inception Institute of Artificial Intelligence
[3] School of Computer Science, Wuhan University [4] Xidian University
(* Equal contributions) http://dpfan.net/CoSOD3k/

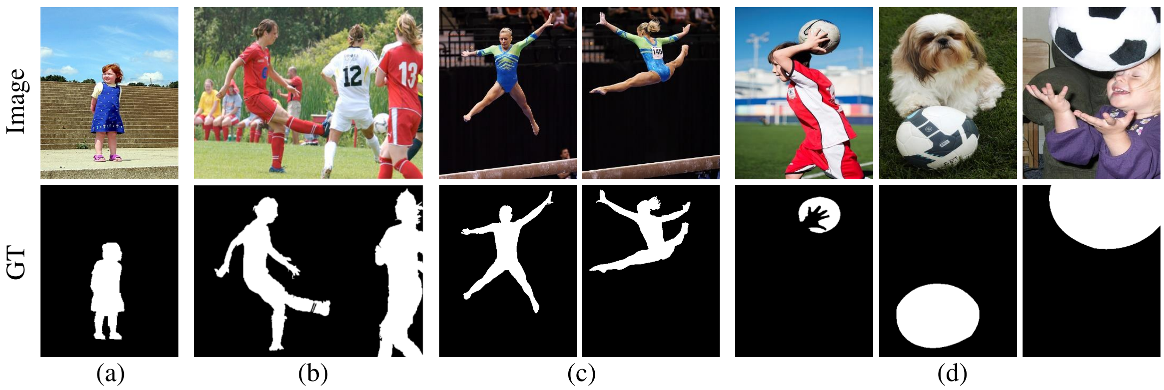

Figure 1: Different salient object detection (SOD) tasks. (a) Traditional SOD [75]. (b) Within-image co-salient object detection (CoSOD) [89], where common salient objects are detected from a single image. (c) Existing CoSOD, where salient objects are detected according to a pair [51] or a group [81] 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.

Abstract

Co-salient object detection (CoSOD) is a newly emerging and rapidly growing branch of salient object detection (SOD), which aims to detect the co-occurring salient objects in multiple images. However, existing CoSOD datasets often have a serious data bias, which assumes that each group of images contains salient objects of similar visual appearances. This bias results in the ideal settings and the effectiveness of the models, trained on existing datasets, may be impaired in real-life situations, where the similarity is usually semantic or conceptual. To tackle this issue, we first collect a new high-quality dataset, named ***CoSOD3k****, which contains 3,316 images divided in 160 groups with multiple level annotations,* i.e.*, category, bounding box, object, and instance levels. CoSOD3k makes a significant leap in terms of diversity, difficulty and scalability, benefiting related vision tasks. Besides, we comprehensively summarize 34 cutting-edge algorithms, benchmarking 19 of them over four existing CoSOD datasets (MSRC, iCoSeg, Image Pair and CoSal2015) and our CoSOD3k with a total of ∼61K images (largest scale), and reporting group-level performance analysis. Finally, we discuss the challenge and future work of CoSOD. Our study would give a strong boost to growth in the CoSOD community. Benchmark toolbox and results are available on our project page.*

1. Introduction

RGB Salient object detection (SOD) [6,18,46,90], RGB-D SOD [22, 25, 98, 103], and Video SOD [23] have been an active [29, 49, 71, 101] research field in computer vision community over the past decade. SOD mimics the human vision system to detect the most attention-grabbing object(s) from individual image, as shown in Fig. 1 (a). As a branch, co-salient object detection (CoSOD) was emerged recently to employ a set of images, which has been attracting growing attention (see Tab. 2) due to its application values in collection-aware crops [34], co-segmentation [77], weakly supervised learning [100], image retrieval [11], image quality assessment [78], and video foreground detection [24], *etc*.

The goal of CoSOD is to extract the salient object(s) that are common among image(s), such as the red-clothed football player or blue-clothed gymnast, in Fig. 1 (b & c). To address this problem, current models tend to focus only on the appearance-similarity between objects. However, this would lead to *data selection bias and is not always appropriate*, since, in real-life applications, salient objects in a group of images often vary in terms of *texture*, *color*, *scene*, and *background* (see our *CoSOD3k* dataset in Fig. 1 (d)), even if they belong to the same category.

To take a deeper look at CoSOD, we make three distinct

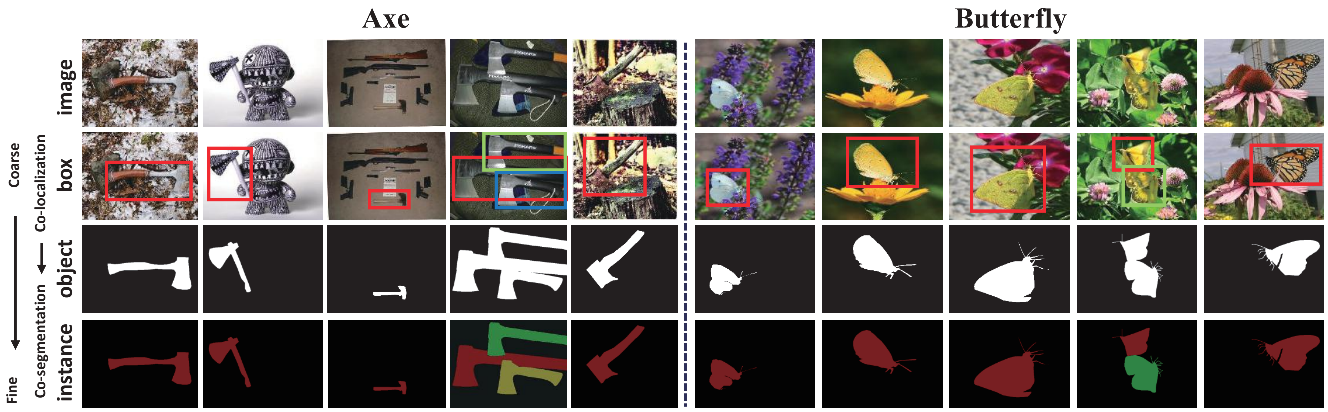

Figure 2: Sample images from our *CoSOD3k* dataset. It has rich annotations, *i.e.*, image-level category (top), bounding boxes, object-level mask, instance-level mask. Our *CoSOD3k* would provide a solid foundation for the CoSOD task and benefit a wide range of related fields, *e.g.*, co-segmentation, weakly supervised localization. Please refer to the supplementary materials for details. Zoom-in for the best view.

contributions:

- First, we construct a challenging *CoSOD3k* dataset, with more realistic settings. Our *CoSOD3k* 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; 2) each image is accompanied by **category**, **bounding box**, **object-level**, and **instance-level** annotations, benefiting various vision tasks, as shown in Fig. 2.
- Second, we present the first large-scale co-salient object detection study, reviewing 34 state-of-the-art (SOTA) models, evaluating 19 of them on four existing CoSOD datasets [4,51,81,93], as well as the proposed *CoSOD3k*. A convenience benchmark toolbox is provided to integrate various publicly available CoSOD datasets and multiple CoSOD metrics to enable convenient performance evaluation.
- Finally, based on our comprehensive evaluation results, we observe several interesting findings and discuss several important issues for future researches. Our research serves as a potential catalyst for promoting large-scale model development and comparison.

2. Related Work

Datasets. Currently, only a few CoSOD datasets have been proposed [4, 11, 51, 81, 89, 93], as shown in Tab. 1. *MSRC* [81] and *Image Pair* [51] 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.* [51], is specially designed for image pairs and contains 210 images (105 groups) in total. The *iCoSeg* [4] 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* [11] and *CoSal2015* [93] are two large-scale publicly available datasets, and the CoSal2015 is widely used for assessing CoSOD algorithms. Different from the above mentioned datasets, the *WICOS* [89] dataset aims to detect co-salient objects from single image, where each image can be viewed as one group.

Dataset	Year	#Gp	#Img	#Avg	IL	Ceg	BBx	HQ	Input
MSRC [81]	2005	8	240	30					Group images
iCoSeg [4]	2010	38	643	17				✓	Group images
Image Pair [51]	2011	105	210	2					Two images
THUR15K [11]	2014	5	15k	3k					Group images
CoSal2015 [93]	2015	50	2,015	40				✓	Group images
WICOS [89]	2018	364	364	1				✓	Single image
CoSOD3k(Ours)	2020	160	3,316	21	✓	✓	✓	✓	Group images

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**: number of average image per group. **HQ**: high-quality annotation. **IL**: whether or not instance-level annotations are provided. **Ceg**: whether or not category labels are provided for each group. **BBx**: whether or not provide bounding box labels are provided for each image.

Although the aforementioned datasets have advanced the CoSOD 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, categories, bounding boxes, instances, *etc.*, which are important for progressing many vision tasks and multi-task modeling.

#	Model	Pub.	Year	#Training	Training Set	Main Component	SL.	Sp.	Po.	Ed.	Post.
1	WPL [34]	UIST	2010			Morphological, Translational Alignment	U				
2	PCSD [10]	ICIP	2010	120,000	8*8 image patch	sparse feature [30], Filter Bank	W				
3	IPCS [51]	TIP	2011			Ncut, co-multilayer Graph	U	✓			
4	CBCS [24]	TIP	2013			Contrast/Spatial/Corresponding Cue	U				
5	MI [50]	TMM	2013			Feature/Images Pyramid, Multi-scale Voting	U	✓			GCut
6	CSHS [59]	SPL	2013			Hierarchical Segmentation, Contour Map [3]	U			✓	
7	ESMG [54]	SPL	2014			Efficient Manifold Ranking [84], OTSU [64]	U				
8	BR [7]	MM	2014			Common/Center Cue, Global Correspondence	U	✓			
9	SACS [8]	TIP	2014			Self-adaptive Weight, Low Rank Matrix	U	✓			
10	DIM‡ [92]	TNNLS	2015	1,000 + 9,963	ASD [1] + PV	SDAE model [92], Contrast/Object Prior	S	✓			
11	CODW‡ [94]	IJCV	2016		ImageNet [16] pre-train	SermaNet [67], RBM [5], IMC, IGS, IGC	W	✓	✓		
12	SP-MIL‡ [96]	TPAMI	2017	(240+643)*10%	MSRC-V1 [81] + iCoseg [4]	SPL [97], SVM, GIST [69], CNNs [9]	W	✓			
13	GD‡ [79]	IJCAI	2017	9,213	MSCOCO [55]	VGGNet16 [68], Group-wise Feature	S				
14	MVSRCC‡ [87]	TIP	2017			LBP, SIFT [61], CH, Bipartite Graph		✓	✓		
15	UMLF [27]	TCSVT	2017	(240 + 2015)*50%	MSRC-V1 [81] + CoSal2015 [94]	SVM, GMR [86], metric learning	S	✓			
16	DML‡ [53]	BMVC	2018	10,000 + 6,232 + 5,168	M10K [12] + THUR-15K [11] + DO	CAE, HSR, Multistage	S				
17	DWSI [89]	AAAI	2018			EdgeBox [106], Low-rank Matrix, CH	S		✓		
18	GONet‡ [33]	ECCV	2018		ImageNet [16] pre-train	ResNet-50 [28], Graphical Optimization	W	✓			CRF
19	COC‡ [31]	IJCAI	2018		ImageNet [16] pre-train	ResNet-50 [28], Co-attention Loss	W		✓		CRF
20	FASS‡ [105]	MM	2018		ImageNet [16] pre-train	DHS [56]/VGGNet, Graph optimization	W	✓			
21	PJO [73]	TIP	2018			Energy Minimization, BoWs	U	✓			
22	SPIG‡ [35]	TIP	2018	10,000+210 +2015+240	M10K [12]+IPCS [51] + CoSal2015 [94] + MSRC-V1 [81]	DeepLab, Graph Representation	S	✓			
23	QGF [36]	TMM	2018		ImageNet [16] pre-train	Dense Correspondence, Quality Measure	S	✓			THR
24	EHL‡ [70]	NC	2019	643	iCoseg [4]	GoogLeNet [72], FSM	S	✓			
25	IML‡ [65]	NC	2019	3624	CoSal2015 [94] + PV + CR	VGGNet16 [68]	S	✓			
26	DGFC‡ [80]	TIP	2019	>200,000	MSCOCO [55]	VGGNet16 [68], Group-wise Feature	S	✓			
27	RCANet‡ [44]	IJCAI	2019	>200,000	MSCOCO [55] + COS + iCoseg [4] + CoSal2015 [94] + MSRC [81]	VGGNet16 [68], Recurrent Units	S				THR
28	GS‡ [74]	AAAI	2019	200,000	COCO-SEG [74]	VGGNet19 [68], Co-category Classification	S				
29	MGCNet‡ [37]	ICME	2019			Graph Convolutional Networks [42]	S	✓			
30	MGLCN‡ [38]	MM	2019	N/A	N/A	VGGNet16, PiCANet [57], Inter-/Intra-graph	S	✓			
31	HC‡ [45]	MM	2019	N/A	N/A	VAE-Net [41], Hierarchical Consistency	S	✓	✓		CRF
32	CSMG‡ [99]	CVPR	2019	25,00	MB [58]	VGGNet16 [68], Shared Superpixel Feature	S	✓			
33	DeepCO3‡ [32]	CVPR	2019	10,000	M10K [12]	SVFSal [95] / VGGNet [68], Co-peak Search	W		✓		
34	GWD‡ [43]	ICCV	2019	>200,000	MSCOCO [55]	VGGNet19 [68], RNN, Group-wise Loss	S				THR

Table 2: Summary of 34 classic and cutting-edge CoSOD approaches. **Training set:** PV = PASCAL VOC07 [17]. CR = Coseg-Rep [15]. DO = DUT-OMRON [86]. 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 model including CBCS [24], RC [12], DCL [49], RFCN [76], DWSI [89]. **SL.** = Supervise 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, GraphCut (GCut), or adaptive/constant threshold (THR), are introduced. ‡ denotes deep models. More details about these models can be found in two survey papers [14, 91].

Traditional Methods. Previous CoSOD studies [8, 27, 51, 73] have found that the inter-image correspondence can be effectively modeled by segmenting the input image into many computational units (*e.g.*, superpixel regions [102], or pixel clusters [24]). A similar observation can be found in recent reviews [14, 91]. In these approaches, heuristic characteristics (*e.g.*, contour [59], 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 [27] or self-adaptive weighting [8]. Several studies have also investigated how to capture inter-image constraints through various computational mechanisms, such as translational alignment [34], efficient manifold ranking [54], and global correspondence [7]. Some methods (*e.g.*, PCSD [10], 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 focused attention occurs.

Deep learning Methods. Deep CoSOD models usually achieve good performance by learning co-salient object representations jointly. More specifically, Zhang *et al.* [92] introduces a domain adaption model to transfer the prior knowledge for CoSOD. Wei *et al.* [79] uses 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 [87] model employed typical features, such as SIFT, LBP and color histograms, as multi-view features. In addition, several other methods [31, 32, 35, 70, 74, 80, 99] are based on the more powerful CNN models (*e.g.*, ResNet [28], Res2Net [26], GoogLeNet [72], VGGNet [68]), achieving SOTA performances. These deep models generally achieved better performance through either weakly-supervised (*e.g.*, CODW [94], SP-MIL [96], GONet [33], FASS [105]) or fully supervised learning (*e.g.*, DIM [92], GD [79], DML [53]). A summary of the traditional and deep learning based models is listed in Tab. 2.

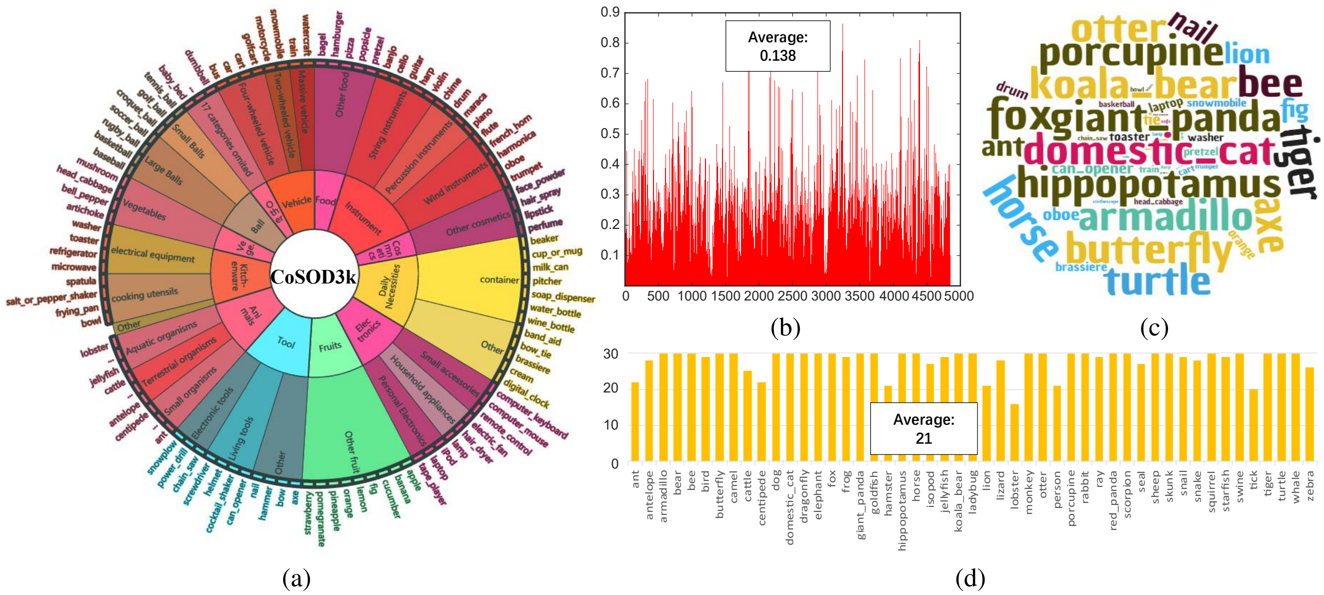

Figure 3: Statistics of the proposed *CoSOD3k* dataset. (a) Taxonomic structure of our dataset. (b) Distribution of the instance sizes. (c) Word clouds of the *CoSOD3k* dataset. (d) Image number of 49 animal categories. **Best viewed on screen and zoomed-in for details.**

3. Proposed *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 [66]. There are several benefits of using ILSVRC to generate our dataset. 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, which provides us a solid basis for building a representative benchmark dataset for CoSOD. More importantly, the accompanying axis-aligned bounding boxes for each target object category allows us to identify unambiguous instance-level annotations.

3.2. Data Annotation

Similar to [21, 63], 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 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-classes*. Finally, we integrate the upper-level class into 13 *super-classes*. The taxonomic structure of our *CoSOD3k* is given in Fig. 3 (a).

Bounding Box Labeling. The second level ananotation is bounding box, 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 [1, 2, 12, 39, 47, 48, 58, 62, 75, 83, 85], we ask three viewers to re-draw the bounding boxes around the object(s) in each image that dominate their attention. Then, we merge the bounding boxes labeled by three viewers and let two additional senior researchers in the CoSOD field double-check the annotations. After that, as done in [40], we discard the images that contain more than six objects, as well as those containing only background. Finally, we collect 3,316 images within 160 categories.

Object-/Instance-level Annotation. The high-quality pixel-level masks are necessary for Co-SOD dataset. 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 to cross-check the whole process by more than three-fold, to ensure high-quality annotation. In this way, we obtain an accurate and challenging dataset with totally 3,316 object-level, and 4,915 instance-level salient object annotations. Note that our final bounding box labels are refined further based on the pixel-level annotation to tighten the target.

3.3. Dataset Features and Statistics

To provide deeper insights into our *CoSOD3k*, we present its several important characteristics in below.

Metric	PCSD [10]	CODR [88]	ESMG [54]	CBCS [24]	IPCS [51]	SACS [8]	UMLF [27]	CSHS [59]	HCNco [60]	DIM [92]‡	EGNet [104]‡	CPD [82]‡	CSMG [99]‡
$S_\alpha \uparrow$	.401	.656	.664	.685	.747	.775	.810	.810	.838	.729	.842	.879	.902
$F_\beta \uparrow$	.378	.652	.651	.800	.786	.837	.870	.856	.867	.867	.835	.880	.925
$E_\xi \uparrow$	.598	.762	.767	.856	.848	.887	.898	.899	.896	.905	.887	.917	.952
$M \downarrow$	.242	.226	.198	.152	.168	.169	.163	.148	.073	.256	.076	.054	.067

Table 3: Benchmarking results of 13 CoSOD approaches on the Image Pair [51] dataset. For simplify, we use $\uparrow$ *and* $\downarrow$ denote larger and smaller is better, respectively. Top three performances are highlighted in red, green and blue.

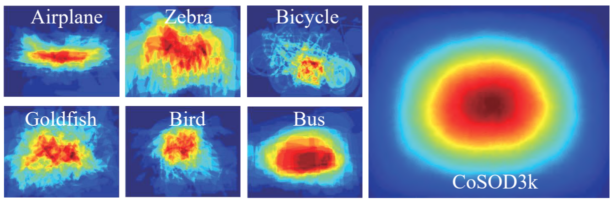

Figure 4: Visualization of overlap masks for mixture-specific category and overall category masks of *CoSOD3k*.

CoSOD3k	Instance Size.			# Instances		
	large (>30%)	middle	small (<5%)	1	2	≥ 3
# Images	439	3173	1303	2371	644	334

Table 4: Statistics of the instance sizes and numbers in the proposed *CoSOD3k* dataset.

Mixture-specific Category Masks. Fig. 4 shows the average ground truth masks for single category and the overall category. It can be observed that some categories with unique shapes (*e.g.*, airplane, zebra, and bicycle) could present the shape-bias maps, while the categories with non-rigid or convex shapes (*e.g.*, goldfish, bird, and bus) may have no clear shape-bias. The overall category mask (the left of Fig. 4) tends to appear a center-bias map without shape bias, which fits the role of salient object. 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 on the supplementary materials.

Sufficient Object Diversity. As shown in Tab. 6 (2^{nd} row) and Fig. 3 (c), our *CoSOD3k* covers a large set of super-classes including *Vegetables*, *Food*, *Fruit*, *Tool*, *Necessary*, *Traffic*, *Cosmetic*, *Ball*, *Instrument*, *Kitchenware*, *Animal* (Fig. 3 d), and *Others*, enabling a comprehensive understanding of real-world scenes.

Size of Instances. The instance size is defined as the ratio of foreground instance pixels to the total image pixels. Tab. 4 summarizes the instance sizes in our *CoSOD3k*. The distributions (Fig. 3 b) of instance sizes are $0.02\% \sim 86.5\%$ (avg.: 13.8%), yielding a broad range.

Number of Instances. Being able to parse object into instance 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 multiple instance scene with instance-level annotation. As reported in Tab. 4, the number of instances (1, 2, $\geq$3) is subject to a ratio of 7:2:1.

4. Benchmark Experiments

4.1. Experimental Settings

Evaluation Metrics. To provide a comprehensive evaluation, two widely-used metrics: maximum F-measure (F_β) [1], MAE (M) [13], and two recently proposed metrics: S-measure (S_α) [19], maximum E-measure (E_ξ) [20] are adapted to evaluating CoSOD performance in multiple images. Let $D = \{G_1, \ldots, G_i, \ldots, G_q\}$ denote the whole dataset with q image groups, and I_k^i is the kth image in image group $G_i = \{I_1^i, \ldots, I_k^i, \ldots, I_{N_i}^i\}$. N_i is the number of images in the G_i. N_D is the total number of images in the whole dataset D. For each metric $\vartheta \in \{S_\alpha, E_\xi, F_\beta, M\}$, we calculate its *mean* score (Tab. 5 & Tab. 3) on the whole dataset. The *mean* metric on dataset D is defined as $Q_\vartheta(D) = \frac{1}{N_D}\sum_{i=1}^{q}\sum_{k=1}^{N_i}\vartheta(I_k^i)$. To provide deep insight into the performance of algorithms on group level, we also provide the *group mean* score, as $T_\vartheta(G_i) = \frac{1}{N_i}\sum_{k=1}^{N_i}\vartheta(I_k^i)$.

Competitors. In this study, we evaluate/compare 19 SOTA CoSOD models, including 10 traditional methods [8, 10, 24, 27, 51, 52, 54, 59, 60, 88] and 9 deep learning models [33, 65, 82, 92, 94, 96, 97, 99, 104]. The methods were chosen based on two criteria: (1) representative, and (2) release code.

Benchmark Protocols. We evaluate on four existing CoSOD datasets, *i.e.*, *Image Pair* [51], *MSRC* [81], *iCoSeg* [4], *CoSal2015* [93], and our *CoSOD3k*. There are 363 groups in total with about 61K images, making this the largest and most comprehensive benchmark. For a fair comparison, we run the available code directly with default settings (*e.g.*, PCSD [10], IPCS [51], CSHS [59], CBCS [24], RFPR [52], ESMG [54], SACS [8], CODR [88], HCNco [60], UMLF [27], CPD [82], EGNet [104]) or using the CoSOD maps provided by the authors (*e.g.*, IML [65], CODW [94], GONet [33], SP-MIL [96], CSMG [99]).

Dataset	Metric	CBCS [24]	ESMG [54]	RFPR [52]	CSHS [59]	SACS [8]	CODR [88]	UMLF [27]	DIM [92]‡	CODW [94]‡	MIL [97]‡	IML [65]‡	GONet [33]‡	SP-MIL [96]‡	CSMG [99]‡	CPD [82]‡	EGNet [104]‡
MSRC	$S_\alpha \uparrow$	.480	.532	.644	.666	.707	.754	.797	.657	.713	.720	.781	.795	.769	.722	.714	.702
	$F_\beta \uparrow$	.630	.606	.696	.727	.782	.776	.849	.705	.784	.768	.840	.846	.824	.847	.762	.752
	$E_\xi \uparrow$	.676	.675	.746	.784	.810	.822	.880	.725	.820	.800	.856	.863	.855	.859	.795	.794
	$M \downarrow$	.314	.303	.302	.289	.224	.198	.184	.309	.264	.216	.174	.179	.218	.190	.173	.186
CoSal2015	$S_\alpha \uparrow$	.544	.552	N/A	.592	.694	.689	.662	.592	.648	.673	-	.751	N/A	.774	.814	.818
	$F_\beta \uparrow$	.532	.476	N/A	.564	.650	.634	.690	.580	.667	.620	-	.740	N/A	.784	.782	.786
	$E_\xi \uparrow$	.656	.640	N/A	.685	.749	.749	.769	.695	.752	.720	-	.805	N/A	.842	.841	.843
	$M \downarrow$	.233	.247	N/A	.313	.194	.204	.271	.312	.274	.210	-	.160	N/A	.130	.098	.099
iCoSeg	$S_\alpha \uparrow$	.658	.728	.744	.750	.752	.815	.703	.758	.750	.727	.832	.820	.771	.821	.861	.875
	$F_\beta \uparrow$	.705	.685	.771	.765	.770	.823	.761	.797	.782	.741	.846	.832	.794	.850	.855	.875
	$E_\xi \uparrow$	.797	.784	.841	.841	.817	.889	.827	.864	.832	.799	.895	.864	.843	.889	.900	.911
	$M \downarrow$	.172	.157	.170	.179	.154	.114	.226	.179	.184	.186	.104	.122	.174	.106	.057	.060

Table 5: Benchmarking results of 16 leading CoSOD approaches on existing three classical [4,81,93] datasets. "N/A" means that the code or results are not available. "–" denotes the whole images of the dataset has been used as training set. Note that the UMLF method adopts half of the images from both MSRC and CoSal2015 to train their model. The "score" indicates the score generated by specific models (*e.g.*, SP-MIL, UMLF) that has been trained on this dataset. Refer to Tab. 2 for more training details (Some methods trained with more data).

4.2. Quantitative Comparisons

Performance on Image Pair. The first CoSOD dataset is the Image Pair [51], as shown in Tab. 3. The Image Pair [51] dataset only has a pair of images in each group, and most co-salient objects have similar appearances. Thus it is relatively easy compared to other co-salient object detection datasets, and the top-1 model, *i.e.*, CSMG [99], gains a high performance (S_α >0.9).

Performance on MSRC. MSRC dataset [81] has more images in each group. From the Tab. 5, it can be observed that *UMLF [27], GONet [33], IML [65], and SP-MIL [96] are the top-4 models on this dataset*. Interestingly, we find that all these models employ the superpixel method to deduce the co-occurrence regions across multiple images. These works obtain good performances on MSRC dataset, which contains a large number of salient objects with similar appearances. However, their performances drop dramatically on iCoSeg (*e.g.*, GONet: No. 2 → No. 5) and our *CoSOD3k* as a consequence of the superpixel technique focusing on color similarity and therefore not being robust enough to semantic-aware datasets.

Performance on iCoSeg. The iCoSeg dataset [4] was originally designed for image co-segmentation but is widely used for the CoSOD task. As can be seen in Tab. 5, the two SOD models (EGNet [104] and CPD [82]) achieve the state-of-the-art performances. One possible reason is that the iCoSeg dataset contains a lot of image with single object, which could be detected easily by SOD model. This partially suggests that iCoSeg dataset may not suit for evaluating co-salient object detection methods.

Performance on CoSal2015. Tab. 5 shows the evaluation results on on the CoSal2015 dataset [93]. One interesting observation is that the top-2 models are still EGNet [104] and CPD [82], which are consistent with the model ranking on the iCoSeg dataset. This implies that some top-performing salient object detection framework may be better suited for extension to CoSOD tasks.

Performance on CoSOD3k. The results on our *CoSOD3k* are presented in Tab. 6. To provide deeper insight into the each group, we report the performances of models on 13 super-classes. We could observe that lower average scores are achieved on classes such as Other (*e.g.*, *baby bed*, *pencil box*), Instrument (*e.g.*, *piano*, *guitar*, *cello*, *etc*), Necessary (*e.g.*, *pitcher*), Tool (*e.g.*, *axe*, *nail*, *chain saw*), and Ball (*e.g.*, *soccer*, *tennis*), which contain complex structures in these real scenes. The top-1 performance (S_α =0.76) of each row clearly shows that the proposed *CoSOD3k* dataset is challenging and leaves abundant room for further research. Note that almost all of the deep-based models (*e.g.*, EGNet [104], CPD [82], IML [65], CSMG [99], *etc*) perform better than the traditional approaches (CODR [88], CSHS [59], CBCS [24], and ESMG [54]), demonstrating the potential advantages in utilizing deep learning techniques to address the CoSOD problem. Another interesting finding is that edge features can help with providing good boundaries for the results. For instance, the best methods from both traditional (CSHS [59]) and deep learning models (*e.g.*, EGNet [104]) introduce edge information to aid detection.

4.3. Qualitative Comparisons

Two visual results of 10 state-of-the-art algorithms on *CoSOD3k* are shown in Fig. 5. It can be seen that the SOD models, *e.g.*, EGNet [104] and CPD [82], detect all salient objects, but ignore the corresponding information. For example, its results of banana contain several other irrelevant objects, *e.g.*, orange, pineapple, and apple. A similar situation also occurs in the images in the horse group, where

	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
CBCS(TIP'13) [24]	.512	.496	.602	.523	.506	.512	.505	.554	.516	.505	.511	.547	.498	.528
CSHS(SPL'13) [59]	.521	.549	.635	.556	.530	.574	.569	.525	.535	.554	.573	.592	.516	.563
ESMG(SPL'14) [54]	.488	.553	.649	.517	.458	.527	.484	.478	.545	.492	.516	.568	.486	.532
CODR(SPL'15) [88]	.632	.646	.696	.595	.586	.649	.602	.574	.576	.612	.616	.682	.573	.630
DIM‡(TNNLS'15) [92]	.593	.626	.663	.538	.534	.569	.530	.515	.540	.528	.545	.577	.517	.559
UMLF(TCSVT'17) [27]	.711	.689	.697	.534	.648	.669	.615	.567	.559	.671	.634	.667	.559	.632
IML‡(NC'19) [65]	.767	.693	.763	.671	.680	.762	.691	.664	.655	.727	.688	.791	.623	.720
CSMG‡(CVPR'19) [99]	.645	.774	.756	.612	.666	.770	.632	.714	.612	.751	.725	.780	.617	.711
CPD‡(CVPR'19) [82]	.769	.732	.788	.705	.733	.824	.719	.676	.611	.796	.745	.846	.649	.757
EGNet‡(ICCV'19) [104]	.795	.746	.792	.712	.740	.809	.728	.683	.621	.800	.742	.850	.659	.762
Average	.643	.650	.704	.596	.608	.667	.608	.595	.577	.644	.630	.690	.570	.639

Table 6: Per super-class average performance (S_α) 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, which release their codes. Note that CPD and EGNet are top-2 SOD models in the socbenchmark (http://dpfan.net/socbenchmark).

the fence (the second image) and the riders (the first and fourth images) are detected together with the horse. On the other hand, the CoSOD methods, *e.g.*, CSMG [99], could identify the common salient objects, but could not produce the accurate predicted map, especially in the object boundaries. Based on the above observations, we conclude that the CoSOD remains far from being solved and there are still large room for the subsequent models.

5. Discussion

From the evaluation, it observes that in most cases, the current SOD methods (*e.g.*, EGNet [104] and CPD [82]) can obtain very competitive or even better performances than the CoSOD methods (*e.g.*, CSMG [99] and SP-MIL [96]). However, this does not mean that the current datasets are not complex enough that directly using the SOD method to obtain good performance—the performances of the SOD methods on the CoSOD datasets are actually lower than those on the SOD datasets, such as HKU-IS [48] (F_β = 0.937 for EGNet) and ECSSD [85] (F_β = 0.943 for EGNet [104]). Instead, this is because many problems in CoSOD are still under-studied, which make the existing CoSOD models less effective. In this section, we discuss four important issues, that have not been fully addressed by the existing co-salient object detection methods and should be studied in the future.

Scalability. The scalability issue is one of the most important issues that need to be considered for designing the CoSOD algorithm. Specifically, it indicates the capability of the CoSOD model for 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 without considering the scalability issue would have huge computational costs and take very long time to run, which are unacceptable in practice. Thus, how to address the scalability issue becomes a key problem in this field, especially when applying CoSOD methods for real-world applications.

Stability. Another important issue is the stability issue. When dealing with image groups containing multiple images, some existing methods (*e.g.*, HCNco [60], PCSD [10], IPCS [51]) divide the image group into image pairs or image sub-groups (*e.g.*, GD [79]). Another school of methods adopt the RNN-based model (*e.g.*,GWD [43]), which need to assign order of the input images. All such strategies would make the whole process unstable as there is no principle ways to divide the image group or assign input order of the related images. This would also influence the application of the CoSOD methods.

Compatibility. Introducing the SOD into the CoSOD is a direct yet effective strategy for building the CoSOD framework. However, the most existing works only introduce the results or features of the SOD model as the useful information cues. One further step for leveraging the SOD technique is to combine the CNN-based SOD network with the CoSOD model to build a unified, end-to-end trainable framework for CoSOD. To achieve this goal, one needs to consider the compatibility of the CoSOD framework, making it convenient to integrate the existing SOD techniques.

Metrics. Current evaluation metrics of CoSOD are designed according to the SOD, *i.e.*, calculating the mean of the SOD scores on each group directly. In contrast to SOD, the CoSOD involves relationship information of co-salient objects among different images, which is more important for CoSOD evaluating and brings more challenges. For example, current CoSOD metrics assume the target objects have the similar sizes in all images. As the objects with different sizes in different images, the CoSOD metric ($S_\alpha, E_\xi, F_\beta, M$ in Sec. 4) would like to be inclined to large objects. Moreover, the current CoSOD metrics are bias to the object detection performance in single image, rather than the identifying of corresponding objects in multiple images. Thus, how to design suitable metrics for CoSOD is an open issue.

Figure 5: Qualitative examples of existing top-10 models on *CoSOD3k*. More examples are shown in the supplementary materials.

6. Conclusion

In this paper, we have presented a complete investigation on the co-salient object detection (CoSOD). By identifying the serious data bias, *i.e.*, assuming that each group of images contain salient object(s) of similar visual appearance, in current CoSOD datasets, we build a new high-quality dataset, named *CoSOD3k*, containing co-salient object(s) that have similarity in semantic or conceptual level. Notably, *CoSOD3k* is the most challenge CoSOD dataset so far, which contains 160 groups and totally 3,316 images annotated with categories, bounding boxes, object-level, and instance-level annotations. It makes a significant leap in terms of diversity, difficulty and scalability, benefiting related vision tasks, *e.g.*, co-segmentation, weakly supervised localization, and instance-level detection, and would benefit a lot for the future development in these research fields.

Besides, this paper has also provided a comprehensive study by summarizing 34 cutting-edge algorithms, benchmarking 19 of them over four existing datasets as well as the proposed *CoSOD3k* dataset. Based on the evaluation results, we provide insightful discussions on the core issues in the research field of CoSOD. We hope the studies presented in this work would give a strong boost to growth in the CoSOD community. In the future, we plan to increase the dataset scale to spark novel ideas.

Acknowledgments. This research was supported by Major Project for New Generation of AI under Grant No. 2018AAA0100400, NSFC (61922046), and Tianjin Natural Science Foundation (17JCJQJC43700).

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *IEEE CVPR*, pages 1597–1604, 2009.

[2] Sharon Alpert, Meirav Galun, Ronen Basri, and Achi Brandt. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *IEEE CVPR*, 2007.

[3] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE TPAMI*, 33(5):898–916, 2010.

[4] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *IEEE CVPR*, 2010.

[5] Yoshua Bengio et al. Learning deep architectures for ai. *FTML*, 2(1):1–127, 2009.

[6] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational Visual Media*, 5(2):117–150, 2019.

[7] Xiaochun Cao, Yupeng Cheng, Zhiqiang Tao, and Huazhu Fu. Co-saliency detection via base reconstruction. In *ACM MM*, pages 997–1000, 2014.

[8] Xiaochun Cao, Zhiqiang Tao, Bao Zhang, Huazhu Fu, and Wei Feng. Self-adaptively weighted co-saliency detection via rank constraint. *IEEE TIP*, 23(9):4175–4186, 2014.

[9] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.

[10] Hwann-Tzong Chen. Preattentive co-saliency detection. In *IEEE ICIP*, pages 1117–1120, 2010.

[11] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu. Salientshape: group saliency in image collections. *The Visual Computer*, 30(4):443–453, 2014.

[12] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015.

[13] Ming-Ming Cheng, Jonathan Warrell, Wen-Yan Lin, Shuai Zheng, Vibhav Vineet, and Nigel Crook. Efficient salient region detection with soft image abstraction. In *IEEE ICCV*, pages 1529–1536, 2013.

[14] Runmin Cong, Jianjun Lei, Huazhu Fu, Ming-Ming Cheng, Weisi Lin, and Qingming Huang. Review of visual saliency detection with comprehensive information. *IEEE TCSVT*, 29(10):2941–2959, 2018.

[15] Jifeng Dai, Ying Nian Wu, Jie Zhou, and Song-Chun Zhu. Cosegmentation and cosketch by unsupervised learning. In *IEEE ICCV*, pages 1305–1312, 2013.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pages 248–255, 2009.

[17] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[18] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, pages 186–202. Springer, 2018.

[19] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A New Way to Evaluate Foreground Maps. In *IEEE ICCV*, pages 4548–4557, 2017.

[20] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *IJCAI*, 2018.

[21] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *IEEE CVPR*, 2020.

[22] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking RGB-D Salient Object Detection: Models, Datasets, and Large-Scale Benchmarks. *IEEE TNNLS*, 2020.

[23] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *IEEE CVPR*, pages 8554–8564, 2019.

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

[25] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. JL-DCF: Joint Learning and Densely-Cooperative Fusion Framework for RGB-D Salient Object Detection. In *IEEE CVPR*, 2020.

[26] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2Net: A New Multi-scale Backbone Architecture. *IEEE TPAMI*, 2020.

[27] Junwei Han, Gong Cheng, Zhenpeng Li, and Dingwen Zhang. A unified metric learning-based framework for co-saliency detection. *IEEE TCSVT*, 28(10):2473–2483, 2017.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016.

[29] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *IEEE CVPR*, pages 1–8, 2007.

[30] Xiaodi Hou and Liqing Zhang. Dynamic visual attention: Searching for coding length increments. In *NIPS*, 2009.

[31] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Co-attention cnns for unsupervised object co-segmentation. In *IJCAI*, pages 748–756, 2018.

[32] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. DeepCO3: Deep Instance Co-Segmentation by Co-Peak Search and Co-Saliency Detection. In *IEEE CVPR*, 2019.

[33] Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, Xiaoning Qian, and Yung-Yu Chuang. Unsupervised CNN-based co-saliency detection with graphical optimization. In *ECCV*, pages 485–501. Springer, 2018.

[34] David E Jacobs, Dan B Goldman, and Eli Shechtman. Cosaliency: Where people look when comparing images. In *ACM UIST*, pages 219–228, 2010.

[35] Dong-ju Jeong, Insung Hwang, and Nam Ik Cho. Co-salient object detection based on deep saliency networks and seed propagation over an integrated graph. *IEEE TIP*, 27(12):5866–5879, 2018.

[36] Koteswar Rao Jerripothula, Jianfei Cai, and Junsong Yuan. Quality-guided fusion-based co-saliency estimation for image co-segmentation and colocalization. *IEEE TMM*, 20(9):2466–2477, 2018.

[37] Bo Jiang, Xingyue Jiang, Jin Tang, Bin Luo, and Shilei Huang. Multiple graph convolutional networks for co-saliency detection. In *IEEE ICME*, pages 332–337, 2019.

[38] Bo Jiang, Xingyue Jiang, Ajian Zhou, Jin Tang, and Bin Luo. A unified multiple graph learning and convolutional network model for co-saliency estimation. In *ACM MM*, pages 1375–1382, 2019.

[39] Huaizu Jiang, Ming-Ming Cheng, Shi-Jie Li, Ali Borji, and Jingdong Wang. Joint Salient Object Detection and Existence Prediction. *Front. Comput. Sci.*, 2017.

[40] Edna L Kaufman, Miles W Lord, Thomas Whelan Reese, and John Volkmann. The discrimination of visual number. *The American Journal of Psychology*, 62(4):498–525, 1949.

[41] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[42] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

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

[44] Bo Li, Zhengxing Sun, Lv Tang, Yunhan Sun, and Jinlong Shi. Detecting robust co-saliency with recurrent co-attention neural network. In *IJCAI*, pages 818–825, 2019.

[45] Bo Li, Zhengxing Sun, Quan Wang, and Qian Li. Co-saliency detection based on hierarchical consistency. In *ACM MM*, pages 1392–1400, 2019.

[46] Chongyi Li, Runmin Cong, Junhui Hou, Sanyi Zhang, Yue Qian, and Sam Kwong. Nested network with two-stream pyramid for salient object detection in optical remote sensing images. *TGRS*, 57(11):9156–9166, 2019.

[47] Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *IEEE CVPR*, pages 247–256, 2017.

[48] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *IEEE CVPR*, 2015.

[49] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *IEEE CVPR*, 2016.

[50] Hongliang Li, Fanman Meng, and King Ngi Ngan. Co-salient object detection from multiple images. *IEEE TMM*, 15(8):1896–1909, 2013.

[51] Hongliang Li and King Ngi Ngan. A co-saliency model of image pairs. *IEEE TIP*, 20(12):3365–3375, 2011.

[52] Lina Li, Zhi Liu, Wenbin Zou, Xiang Zhang, and Olivier Le Meur. Co-saliency detection based on region-level fusion and pixel-level refinement. In *IEEE ICME*, 2014.

[53] Min Li, Shizhong Dong, Kun Zhang, Zhifan Gao, Xi Wu, Heye Zhang, Guang Yang, and Shuo Li. Deep learning intra-image and inter-images features for co-saliency detection. In *BMVC*, page 291, 2018.

[54] Yijun Li, Keren Fu, Zhi Liu, and Jie Yang. Efficient saliency-model-guided visual co-saliency detection. *IEEE SPL*, 22(5):588–592, 2014.

[55] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[56] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *IEEE CVPR*, pages 678–686, 2016.

[57] Nian Liu, Junwei Han, and Ming-Hsuan Yang. PiCANet: Learning pixel-wise contextual attention for saliency detection. In *IEEE CVPR*, pages 3089–3098, 2018.

[58] Tie Liu, Jian Sun, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. In *IEEE CVPR*, pages 1–8, 2007.

[59] Zhi Liu, Wenbin Zou, Lina Li, Liquan Shen, and Olivier Le Meur. Co-saliency detection based on hierarchical segmentation. *IEEE SPL*, 21(1):88–92, 2013.

[60] Jing Lou, Fenglei Xu, Qingyuan Xia, Wankou Yang, and Mingwu Ren. Hierarchical co-salient object detection via color names. In *IEEE ACPR*, pages 718–724, 2017.

[61] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[62] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *IEEE ICCV*, 2001.

[63] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *CVPR*, pages 909–918, 2019.

[64] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE TSMC*, 9(1):62–66, 1979.

[65] Jingru Ren, Zhi Liu, Xiaofei Zhou, Cong Bai, and Guangling Sun. Co-saliency detection via integration of multi-layer convolutional features and inter-image propagation. *Neurocomputing*, 371:137–146, 2020.

[66] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[67] Pierre Sermanet, David Eigen, Xiang Zhang, Micha`el Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014.

[68] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[69] Parthipan Siva, Chris Russell, Tao Xiang, and Lourdes Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *IEEE CVPR*, pages 3238–3245, 2013.

[70] Shaoyue Song, Hongkai Yu, Zhenjiang Miao, Dazhou Guo, Wei Ke, Cong Ma, and Song Wang. An easy-to-hard learning strategy for within-image co-saliency detection. *Neurocomputing*, 358:166–176, 2019.

[71] Jinming Su, Jia Li, Yu Zhang, Changqun Xia, and Yonghong Tian. Selectivity or invariance: Boundary-aware salient object detection. In *IEEE ICCV*, 2019.

[72] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE CVPR*, pages 1–9, 2015.

[73] Chung-Chi Tsai, Weizhi Li, Kuang-Jui Hsu, Xiaoning Qian, and Yen-Yu Lin. Image co-saliency detection and co-segmentation via progressive joint optimization. *IEEE TIP*, 28(1):56–71, 2018.

[74] Chong Wang, Zheng-Jun Zha, Dong Liu, and Hongtao Xie. Robust deep co-saliency detection with group semantic. In *AAAI*, pages 8917–8924, 2019.

[75] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *IEEE CVPR*, pages 136–145, 2017.

[76] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Saliency detection with recurrent fully convolutional networks. In *ECCV*, pages 825–841, 2016.

[77] Wenguan Wang and Jianbing Shen. Higher-order image co-segmentation. *IEEE TMM*, 18(6):1011–1021, 2016.

[78] Xiaochuan Wang, Xiaohui Liang, Bailin Yang, and Frederick WB Li. No-reference synthetic image quality assessment with convolutional neural network and local image saliency. *Computational Visual Media*, 2019.

[79] Lina Wei, Shanshan Zhao, Omar El Farouk Bourahla, Xi Li, and Fei Wu. Group-wise deep co-saliency detection. In *IJCAI*, 2017.

[80] Lina Wei, Shanshan Zhao, Omar El Farouk Bourahla, Xi Li, Fei Wu, and Yueting Zhuang. Deep group-wise fully convolutional network for co-saliency detection with graph propagation. *IEEE TIP*, 28(10):5052–5063, 2019.

[81] John Winn, Antonio Criminisi, and Tom Minka. Object categorization by learned universal visual dictionary. In *IEEE ICCV*, pages 1800–1807, 2005.

[82] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *IEEE CVPR*, pages 3907–3916, 2019.

[83] Changqun Xia, Jia Li, Xiaowu Chen, Anlin Zheng, and Yu Zhang. What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors. In *IEEE CVPR*, pages 4142–4150, 2017.

[84] Bin Xu, Jiajun Bu, Chun Chen, Deng Cai, Xiaofei He, Wei Liu, and Jiebo Luo. Efficient manifold ranking for image retrieval. In *ACM SIGIR*, pages 525–534, 2011.

[85] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *IEEE CVPR*, pages 1155–1162, 2013.

[86] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *IEEE CVPR*, pages 3166–3173, 2013.

[87] Xiwen Yao, Junwei Han, Dingwen Zhang, and Feiping Nie. Revisiting co-saliency detection: A novel approach based on two-stage multi-view spectral rotation co-clustering. *IEEE TIP*, 26(7):3196–3209, 2017.

[88] Linwei Ye, Zhi Liu, Junhao Li, Wan-Lei Zhao, and Liquan Shen. Co-saliency detection via co-salient object discovery and recovery. *IEEE SPL*, 22(11):2073–2077, 2015.

[89] Hongkai Yu, Kang Zheng, Jianwu Fang, Hao Guo, Wei Feng, and Song Wang. Co-saliency detection within a single image. In *AAAI*, 2018.

[90] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *IEEE ICCV*, pages 7234–7243, 2019.

[91] Dingwen Zhang, Huazhu Fu, Junwei Han, Ali Borji, and Xuelong Li. A review of co-saliency detection algorithms: Fundamentals, applications, and challenges. *ACM TIST*, 9(4):1–31, 2018.

[92] Dingwen Zhang, Junwei Han, Jungong Han, and Ling Shao. Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining. *IEEE TNNLS*, 27(6):1163–1176, 2015.

[93] Dingwen Zhang, Junwei Han, Chao Li, and Jingdong Wang. Co-saliency detection via looking deep and wide. In *IEEE CVPR*, pages 2994–3002, 2015.

[94] Dingwen Zhang, Junwei Han, Chao Li, Jingdong Wang, and Xuelong Li. Detection of co-salient objects by looking deep and wide. *IJCV*, 120(2):215–232, 2016.

[95] Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *IEEE ICCV*, pages 4048–4056, 2017.

[96] Dingwen Zhang, Deyu Meng, and Junwei Han. Co-saliency detection via a self-paced multiple-instance learning framework. *IEEE TPAMI*, 39(5):865–878, 2016.

[97] Dingwen Zhang, Deyu Meng, Chao Li, Lu Jiang, Qian Zhao, and Junwei Han. A self-paced multiple-instance learning framework for co-saliency detection. In *IEEE ICCV*, pages 594–602, 2015.

[98] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. UC-Net: Uncertainty Inspired RGB-D Saliency Detection via Conditional Variational Autoencoders. In *IEEE CVPR*, 2020.

[99] Kaihua Zhang, Tengpeng Li, Bo Liu, and Qingshan Liu. Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing. In *CVPR*, pages 3095–3104, 2019.

[100] Lu Zhang, Jianming Zhang, Zhe Lin, Huchuan Lu, and You He. Capsal: Leveraging captioning to boost semantics for salient object detection. In *IEEE CVPR*, 2019.

[101] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *IEEE ICCV*, pages 202–211, 2017.

[102] Jiaxing Zhao, Ren Bo, Qibin Hou, Ming-Ming Cheng, and Paul Rosin. Flic: Fast linear iterative clustering with active search. *Computational Visual Media*, 4(4):333–348, 2018.

[103] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and fluid pyramid integration for rgbd salient object detection. In *IEEE CVPR*, pages 3927–3936, 2019.

[104] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. EGNet: Edge Guidance Network for Salient Object Detection. In *IEEE ICCV*, pages 8779–8788, 2019.

[105] Xiaoju Zheng, Zheng-Jun Zha, and Liansheng Zhuang. A feature-adaptive semi-supervised framework for co-saliency detection. In *ACM MM*, pages 959–966, 2018.

[106] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014.