Taking a Deeper Look at Co-Salient Object Detection

Deng-Ping Fan\(^1,2,\ast\) Zheng Lin\(^1,\ast\) Ge-Peng Ji\(^3\) Dingwen Zhang\(^4\) Huazhu Fu\(^2\) Ming-Ming Cheng\(^1\)\(^\otimes\)

\(^1\) TKLNDST, College of CS, Nankai University \(^2\) Inception Institute of Artificial Intelligence \(^3\) School of Computer Science, Wuhan University \(^4\) Xidian University

\((\ast\text{ Equal contributions})\) 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 \(~\text{61K}\) 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
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.

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.
## 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-adjoint feature 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.

### Table 2: Summary of 34 classic and cutting-edge CoSOD approaches.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WPL [34]</td>
<td>2011</td>
<td>120,000</td>
<td>8*8 image patch</td>
<td>Motophotometric, Translational Alignment</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>PCSD [10]</td>
<td>2010</td>
<td>120,000</td>
<td>Ncut, co-multilayer Graph</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>IPCS [51]</td>
<td>2011</td>
<td>120,000</td>
<td>Contrast/Spatial/Corresponding Cue</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>CBCS [24]</td>
<td>2011</td>
<td>120,000</td>
<td>Feature/Imagery Pyramid, Multi-scale Voting</td>
<td>✓</td>
<td>GCut</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>TMM [50]</td>
<td>2011</td>
<td>120,000</td>
<td>Hierarchical Segmentation, Contour Map [3]</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>CSHS [59]</td>
<td>2011</td>
<td>120,000</td>
<td>Efficient Manifold Ranking [84], OTSU [64]</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>ESFG [54]</td>
<td>2011</td>
<td>120,000</td>
<td>Common/Center Cue, Global Correspondence</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>BR [7]</td>
<td>2011</td>
<td>120,000</td>
<td>Self-adaptive Weight, Low Rank Matrix</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>SACS [8]</td>
<td>2011</td>
<td>120,000</td>
<td>SVM, GMR [24]</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>DIM [12]</td>
<td>2011</td>
<td>120,000</td>
<td>SVM, GMR [24]</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>CODW [94]</td>
<td>2011</td>
<td>120,000</td>
<td>SVM, GMR [24]</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>SP-MIL [96]</td>
<td>2011</td>
<td>120,000</td>
<td>SVM, GMR [24]</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>GD [79]</td>
<td>2011</td>
<td>120,000</td>
<td>SVM, GMR [24]</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>MVSRCC [87]</td>
<td>2011</td>
<td>120,000</td>
<td>SVM, GMR [24]</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

Denotes deep models. More details about these models can be found in two survey papers [14, 91].
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.

![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.](image-url)
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 (2nd row) and Fig. 3 (c), our CosOD3k covers a large set of superclasses 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 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 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 ($F_\beta$) [1], MAE ($M$) [13], and two recently proposed metrics: S-measure ($S_\alpha$) [19], maximum E-measure ($E_\xi$) [20] are adapted to evaluating CosOD performance in multiple images. Let $D = \{G_1, \ldots, G_q\}$ denote the whole dataset with $q$ image groups, and $I_{ik}$ is the $k$th image in image group $G_i = \{I_{i1}, \ldots, I_{ik}, \ldots, I_{iN_i}\}$. $N_i$ 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 $Q_\vartheta(D) = \frac{1}{N_D} \sum_{i=1}^{N_q} \sum_{k=1}^{N_i} \vartheta(I_{ik})$.

**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], HC-NoCo [60], UMLF [27], CPD [82], EGONet [104]) or using the CoSOD maps provided by the authors (e.g., IML [65], CODKW [94], GONet [33], SP-MIL [96], CSMG [99]).
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 ($S_\alpha > 0.9$).

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 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 ($S_\alpha = 0.76$) 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
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CSHS/SPL'13 [59]</td>
<td>.521</td>
<td>.549</td>
<td>.635</td>
<td>.556</td>
<td>.530</td>
<td>.574</td>
<td>.569</td>
<td>.525</td>
<td>.535</td>
<td>.554</td>
<td>.573</td>
<td>.592</td>
<td>.516</td>
<td>.563</td>
</tr>
<tr>
<td>ESMG/SPL'14 [54]</td>
<td>.488</td>
<td>.553</td>
<td>.649</td>
<td>.517</td>
<td>.458</td>
<td>.527</td>
<td>.484</td>
<td>.478</td>
<td>.545</td>
<td>.492</td>
<td>.516</td>
<td>.568</td>
<td>.486</td>
<td>.532</td>
</tr>
<tr>
<td>CODR/SPL'15 [88]</td>
<td>.632</td>
<td>.646</td>
<td>.696</td>
<td>.595</td>
<td>.586</td>
<td>.649</td>
<td>.602</td>
<td>.574</td>
<td>.576</td>
<td>.612</td>
<td>.616</td>
<td>.682</td>
<td>.573</td>
<td>.630</td>
</tr>
<tr>
<td>DCM/TNNLS'15 [92]</td>
<td>.593</td>
<td>.626</td>
<td>.663</td>
<td>.538</td>
<td>.534</td>
<td>.569</td>
<td>.530</td>
<td>.515</td>
<td>.540</td>
<td>.528</td>
<td>.545</td>
<td>.577</td>
<td>.517</td>
<td>.559</td>
</tr>
<tr>
<td>UMLF/TCSVT'17 [27]</td>
<td>.711</td>
<td>.689</td>
<td>.697</td>
<td>.594</td>
<td>.648</td>
<td>.669</td>
<td>.615</td>
<td>.567</td>
<td>.559</td>
<td>.671</td>
<td>.634</td>
<td>.667</td>
<td>.559</td>
<td>.632</td>
</tr>
<tr>
<td>IML[NC'19] [65]</td>
<td>.767</td>
<td>.693</td>
<td>.763</td>
<td>.671</td>
<td>.680</td>
<td>.762</td>
<td>.691</td>
<td>.664</td>
<td>.655</td>
<td>.727</td>
<td>.688</td>
<td>.791</td>
<td>.623</td>
<td>.720</td>
</tr>
<tr>
<td>CSMD[CVPR'19] [99]</td>
<td>.645</td>
<td>.774</td>
<td>.756</td>
<td>.612</td>
<td>.666</td>
<td>.770</td>
<td>.632</td>
<td>.714</td>
<td>.612</td>
<td>.751</td>
<td>.725</td>
<td>.780</td>
<td>.617</td>
<td>.711</td>
</tr>
<tr>
<td>CPD[TIP'19] [82]</td>
<td>.769</td>
<td>.763</td>
<td>.788</td>
<td>.705</td>
<td>.733</td>
<td>.824</td>
<td>.719</td>
<td>.676</td>
<td>.611</td>
<td>.796</td>
<td>.745</td>
<td>.846</td>
<td>.649</td>
<td>.757</td>
</tr>
<tr>
<td>EGN[ICCV'19] [104]</td>
<td>.795</td>
<td>.746</td>
<td>.792</td>
<td>.712</td>
<td>.740</td>
<td>.809</td>
<td>.728</td>
<td>.683</td>
<td>.621</td>
<td>.800</td>
<td>.742</td>
<td>.858</td>
<td>.659</td>
<td>.762</td>
</tr>
</tbody>
</table>


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] ($F_\beta = 0.937$ for EGNet) and ECSSD [85] ($F_\beta = 0.943$ 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], FCSOD [10], IPCS [51]) divide the image group into image pairs or image sub-groups (e.g., GD [79]). Another school of methods adopts 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 biased 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.
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, benefitting 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 (17JCQJC43700).
References


