Democracy Does Matter: Comprehensive Feature Mining for Co-Salient Object Detection

Siyue Yu[1,2],   Jimin Xiao[1]*   Bingfeng Zhang[1,2],   Eng Gee Lim[1]

[1]XJTLU,   [2]University of Liverpool

{siyue.yu,jimin.xiao,bingfeng.zhang,enggee.lim}@xjtlu.edu.cn

Abstract

Co-salient object detection, with the target of detecting co-existed salient objects among a group of images, is gaining popularity. Recent works use the attention mechanism or extra information to aggregate common co-salient features, leading to incomplete even incorrect responses for target objects. In this paper, we aim to mine comprehensive co-salient features with democracy and reduce background interference without introducing any extra information. To achieve this, we design a democratic prototype generation module to generate democratic response maps, covering sufficient co-salient regions and thereby involving more shared attributes of co-salient objects. Then a comprehensive prototype based on the response maps can be generated as a guide for final prediction. To suppress the noisy background information in the prototype, we propose a self-contrastive learning module, where both positive and negative pairs are formed without relying on additional classification information. Besides, we also design a democratic feature enhancement module to further strengthen the co-salient features by readjusting attention values. Extensive experiments show that our model obtains better performance than previous state-of-the-art methods, especially on challenging real-world cases (e.g., for CoCA, we obtain a gain of 2.0% for MAE, 5.4% for maximum F-measure, 2.3% for maximum E-measure, and 3.7% for S-measure) under the same settings. Source code is available at https://github.com/siyueyu/DCFM.

1. Introduction

Co-salient object detection (CoSOD) aims to detect the common salient objects among a group of input images. Unlike salient object detection (SOD), which is to detect the most attractive objects by mimicking human eyes [3,21,

'Butterfly'

Figure 1. Visualization of response maps. (a) Inputs; (b) Response maps generated by the previous approach [11]; (c) Ours. It can be seen that ours can cover more co-salient objects.

24, 27, 30, 31, 36, 45], CoSOD focuses on detecting salient yet co-existed objects among all the input images. In this case, CoSOD faces two main challenges: 1) reduce the interference of noisy background in complex scenes; 2) mine integral co-salient objects with large appearance variations. Some works introduce extra SOD dataset to provide saliency guidance [52, 53] or predict saliency maps [19] in order to mask out the co-salient objects. However, these approaches highly depend on the extra dataset, leading to supererogatory human effort to provide annotations.

Recent approaches [11, 19, 49, 52] try to use attention mechanism [39] to strengthen co-salient features or build feature consistency to formulate the shared attributes of co-salient objects for integral predictions. However, there are two main drawbacks when directly applying attention mechanism for this task. On the one hand, the response maps reflecting the shared attributes, obtained in the attention mechanism, can only cover limited pixels belonging to co-salient objects, as shown in Fig. 1.(b). In this case, it is difficult for the model to learn comprehensive shared attributes of co-salient objects. On the other hand, for complex scenes, the attention mechanism tends to focus on the

wrong object regions, as shown in the second picture of Fig. 1.(b). Some methods such as GCoNet [11] propose a kind of group collaborative learning by collecting artificial negative group pairs. However, their pairs are grouped based on the auxiliary classification information, which requires major effort to group dissimilar negative category pairs as there is no clear definition of natural discrete object categories in real world [37].

To solve aforementioned issues, we design a novel **D**emocratic **C**o-salient-**F**eature-**M**ining framework (**DCFM**). Our DCFM can directly mine more comprehensive features and suppress the noisy background effectively without using extra SOD dataset or classification information. Specifically, in order to mine sufficient co-salient information, we first design a democratic prototype generation module (DPG), where democratic response maps are generated to capture more shared attributes. As shown in Fig. 1.(c), our response maps cover more regions of co-salient objects. Then, a prototype with comprehensive co-salient information can be generated according to the democratic response maps, which can further guide the model to predict the co-salient objects.

Next, in order to suppress noisy background information in our prototype and avoid introducing extra classification information, we propose a simple self-contrastive learning module (SCL) to form positive and negative pairs to filter noise. We argue that the prototype generated from original images should be consistent with that generated when the image background regions are erased, and should be different from that generated when the co-salient objects are erased. Thus, a self-contrastive loss among these prototypes is designed to suppress the influence of noisy background and help the model learn more discriminative features of co-salient objects.

Finally, to further strengthen the detected co-salient features from the above modules, we design a democratic feature enhancement module (DFE) based on the attention mechanism [39]. As mentioned before, the attention mechanism tends to focus on a limited number of correlated features, which fails to provide comprehensive information. Therefore, we readjust the attention values to generate a democratic attention map aggregating more correlated pixels for feature enhancement.

Generally, our main contributions can be summarized as:

- A democratic prototype generation module (DPG) is designed to build response maps covering sufficient co-salient regions, so as to generate a prototype containing comprehensive shared attributes as guidance for co-saliency prediction.

- A self-contrastive learning module (SCL) is proposed to help our model reduce the influence of noisy background without relying on additional classification in-

formation, where both positive and negative samples are generated from the image itself.

- A democratic feature enhancement module (DFE) is designed to further strengthen the co-salient features by adjusting attention values to involve more related pixels.

- Extensive experiments show that our method performs better than state-of-the-art methods, especially on challenging real-world cases, such as the CoCA dataset, we obtain a gain of 2.0% for MAE, 5.4% for maximum F-measure, 2.3% for maximum E-measure, and 3.7% for S-measure under the same settings.

2. Related Work

2.1. Co-Salient Object Detection

CoSOD is gaining increasing popularity. Some works establish graphs to model the relationship among pixels from a group of images [15–18, 43, 51], then the co-salient objects can be mined with consistent features. Some works adopt extra salient object detection to mine salient objects first and then conduct CoSOD [19, 52, 53]. Besides, SAEF [38] proposes to use saliency proposals generated by unsupervised deep learning based models first and then conduct CoSOD according to those proposals. Other works [11, 49, 54] try to formulate shared attributes among input images to reflect the co-salient pixels and use classification information as a supplement of semantic information. In CoEGNet [9], edge detection is used for better structure prediction. More information on CoSOD can be found in surveys [5, 10, 47]. Although these methods have obtained outstanding performances, they rely on extra information to learn discriminative co-salient features. Therefore, we consider thoroughly exploring the intrinsic characteristics of co-salient objects and background to realize CoSOD without using the SOD dataset or extra classification information.

2.2. Contrastive Learning

Contrastive learning is prevalent in self-supervised classification. SimCLR [2] studies the importance of grouping negative pairs for contrastive learning. MoCo [14] adopts a memory bank to reduce the influence of insufficient negative pairs. Additionally, contrastive learning has also be applied in many other tasks, such as video grounding [28], long-tailed recognition [6], action recognition [34] and visual localization [37]. However, as argued in [37], it is difficult to define artificial negative pairs for contrastive learning, and they use soft assignments of images instead. In this paper, we also study how to group positive/negative pairs without artificial defined categories in CoSOD.

2.3. Attention Mechanism

Attention mechanism has been applied in different tasks. It has been applied in machine translation to draw global relationships between input and output [39]. It is also utilized in non-local networks to deal with detection and segmentation [12,41,42]. Besides, A^2GNN [46] utilize it for affinity maps in weakly semantic segmentation. Additionally, it is used in video object segmentation to segment and track the target object [29, 44]. Moreover, it is deployed in referring expression grounding [35] and language-person search [23]. Recently, attention mechanism has been applied in SOD or CoSOD. MSANet [55] applies attention mechanism to mine salient features and suppress background information. Besides, it is also used in CA-FCN [13] and RCAU [22] as co-attention to link co-salient objects for CoSOD. However, we find that the attention mechanism tends to focus on a limited number of pixels. Therefore, in this paper, we try to introduce democracy into the attention mechanism to involve more related pixels.

3. Methodology

3.1. Overview

The CoSOD dataset includes groups of images with labels. Each group is represented as $G = \{I, Y\}$, where $I = \{x_n\}_{n=1}^N$, $Y = \{y_n\}_{n=1}^N$, x_n is the input image, y_n is the corresponding label, N is the total number of images in group G, and all images contain related objects. The labels are unavailable during inference. The model needs to detect the co-existed salient objects in each image of the same group. In this work, we aim to design a model that can detect the co-salient objects by thoroughly exploring the shared attributes to mine comprehensive co-salient features, and suppress noisy background through self-contrastive learning without using classification information or extra SOD dataset.

The framework of our method and the learning procedure are demonstrated in Fig. 2. There are five main modules in our network, including a feature extractor, a democratic prototype generation module (DPG), a self-contrastive learning module (SCL), a democratic feature enhancement module (DFE), and a decoder. Note that the SCL is only applied for training and will be removed during inference. The overall process can be summarized as:

1) Firstly, the feature extractor encodes a group of relative images (N images) as initial features, which are then proceeded by the DPG to generate a comprehensive co-salient prototype.

2) Meanwhile, to avoid mining noisy information from the background in the prototype, our SCL is deployed for auxiliary training.

3) Then, the prototype is fused into the visual features, and the fused features are transmitted into the DFE to strengthen the features further.

4) Finally, the strengthened features are input into the decoder to predict the corresponding co-saliency maps.

In the following sections, the details about the democratic prototype generation module, the self-contrastive learning module, and the democratic feature enhancement module will be discussed, respectively.

3.2. Democratic Prototype Generation Module

Our democratic prototype generation module (DPG) mainly contains three parts in series, which are the residual block, the seed selection block (SSB), and the democratic response block (DRB).

After passing the feature extractor, we obtain the initial features $\mathcal{F}_{ext} \in \mathbb{R}^{N \times C \times H \times W}$ (C, H, W are the channel number, height, and width), which are processed by the residual block first to generate strengthened residual features $\mathcal{F}_{res}$:

$$\mathcal{F}_{res} = \mathcal{F}_{ext} + conv^{1 \times 1}(\mathcal{F}_{ext}), \tag{1}$$

where $conv^{1 \times 1}$ represents for the 1×1 convolution layer and $\mathcal{F}_{res} \in \mathbb{R}^{N \times C \times H \times W}$.

Then, the generated features $\mathcal{F}_{res}$ are passed into the SSB to select the most discriminative seeds for the co-salient objects in each input image. Next, the selected seeds are correlated with the residual feature maps to produce the response maps by the DRB. Finally, the response maps are multiplied with the residual features and averaged to generate the prototype, containing comprehensive co-salient feature information and guiding following prediction.

Seed Selection Block (SSB). The SSB is demonstrated in Fig. 3. This block is deployed to detect each image's most representative pixel as seed for response map generation. First, the residual features $\mathcal{F}_{res}$ are input to our SSB. Then, the attention mechanism is employed, in which two 1×1 convolution layers are deployed to obtain two feature maps, namely $K \in \mathbb{R}^{N \times C \times H \times W}$ and $Q \in \mathbb{R}^{N \times C \times H \times W}$. After reshaping both K and Q to shape $\mathbb{R}^{NHW \times C}$, the feature similarity map (S) of each pixel is computed as

$$S = KQ^\top, \tag{2}$$

where $S \in \mathbb{R}^{NHW \times NHW}$, $\top$ means transpose, and each row of S represents similarities between one pixel and all pixels of the N inputs.

Then, we first reshape S into $S \in \mathbb{R}^{NHW \times N \times HW}$ and choose its maximum similarity value in each image, to get N maximum similarity values for each pixel. This process is calculated by

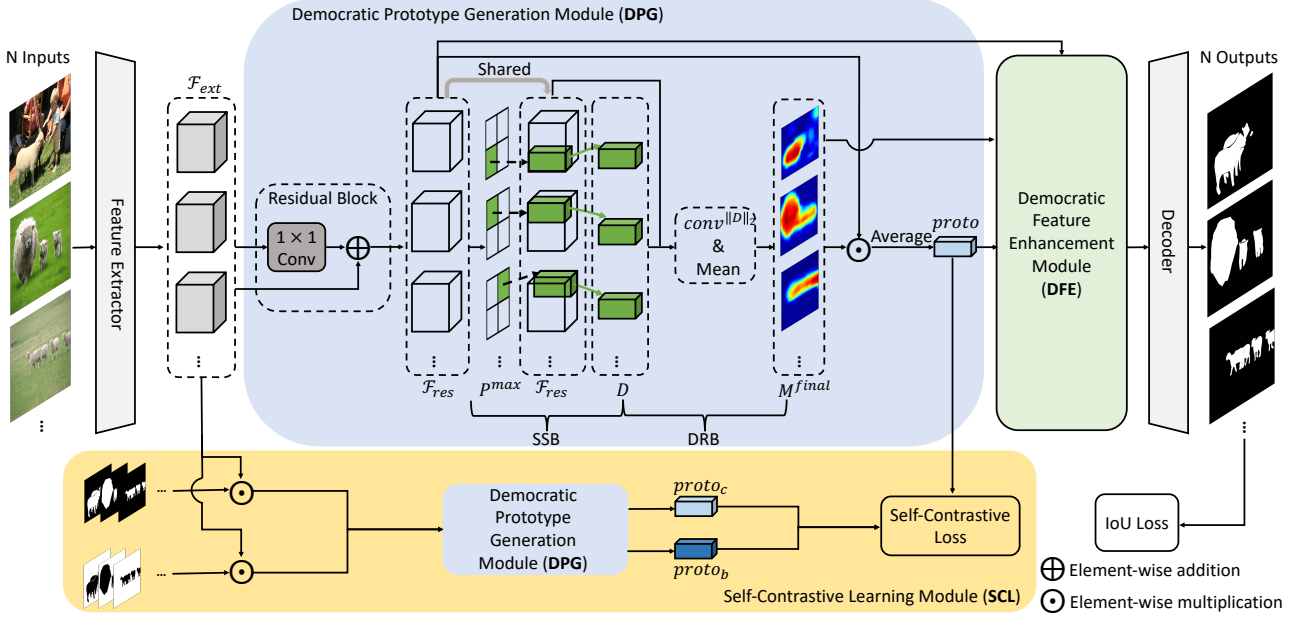$$S^{N\text{-}max} = \max_{i=1 \cdots HW} S[:, :, i], \tag{3}$$

Figure 2. The framework of our network and the learning procedure. Specifically, the network contains five main parts, including a feature extractor, a democratic prototype generation module (DPG), a self-contrastive learning module (SCL), a democratic feature enhancement module (DFE), and a decoder. Note that the SCL is only used during training.

where $S^{N\text{-}max} \in \mathbb{R}^{NHW \times N}$. Afterwards, the average of the N maximum similarity values is treated as the co-salient probability of each pixel,

$$P = \frac{1}{N} \sum_{n=1}^{N} S^{N\text{-}max}[:, n], \qquad (4)$$

where $P \in \mathbb{R}^{NHW}$.

Then, the probability map is reshaped back to $P \in \mathbb{R}^{N \times H \times W}$. We can locate the pixel with the highest probability of being the co-salient object in each image by

$$P^{max} = \max_{\substack{h=1,\cdots H \\ w=1,\cdots W}} P[:, h, w], \qquad (5)$$

$$index = ind(P^{max}), \qquad (6)$$

where $ind(\cdot)$ means taking out the index of P^{max}.

Finally, we take out the feature vectors from the $\mathcal{F}_{res}$ according to the index in Eq. (6) as the final seeds by

$$D = \mathcal{F}_{res}(index). \qquad (7)$$

Note that each image will provide one seed vector, and there are totally N seeds. These seeds can represent the essential characteristics of the co-salient objects in each input image and be used for localization.

Democratic Response Block (DRB). The DRB is demonstrated in Fig. 3. If we directly use the seeds D as

the prototype, it fails to aggregate comprehensive characteristics of the co-salient objects. This is because it is difficult for limited seeds to express the integral co-existed objects, especially when there are large appearance variations among the group. Thus, we try to involve more pixels of the co-salient objects to generate a comprehensive prototype by considering the correlation between each pixel and the seeds D from SSB.

Specifically, we first use L2 normalization in channel dimension to obtain the normalized residual features $\|\mathcal{F}_{res}\|_2$ and the normalized seeds $\|D\|_2$. Then, $\|D\|_2$ are treated as the kernel to conduct convolution on the $\|\mathcal{F}_{res}\|_2$:

$$M = conv^{\|D\|_2}(\|\mathcal{F}_{res}\|_2), \qquad (8)$$

where M means response maps, $conv^{\|D\|_2}$ is the convolution with $\|D\|_2$ as kernel. Since D has N seed vectors, the size of response maps become $\mathbb{R}^{N \times N \times H \times W}$ after Eq. (8), with channel dimension being the number of response maps for each input.

The final democratic response map of each image is computed as the mean value of the N response maps:

$$M^{final} = \frac{1}{N} \sum_{n=1}^{N} M[:, n, :, :], \qquad (9)$$

where $M^{final} \in \mathbb{R}^{N \times H \times W}$. In this way, more pixels have chance to contribute to the response maps.

Finally, the prototype ($proto \in \mathbb{R}^{1 \times C}$) is generated by

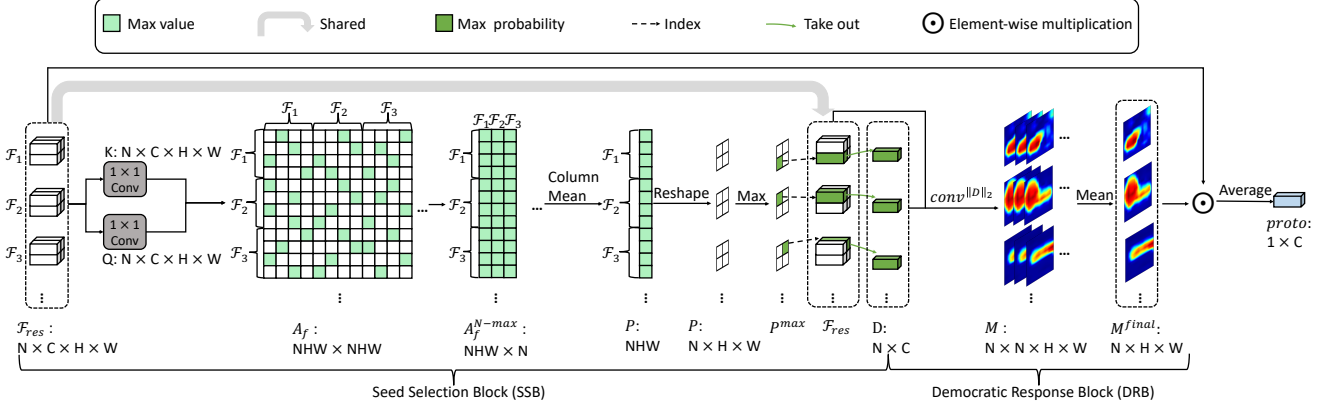$$proto = avg(M^{final} \odot \mathcal{F}_{res}), \qquad (10)$$

Figure 3. The framework of the seed selection block (SSB) and democratic response block (DRB). The inputs are the residual features. Then, the co-salient seeds are selected first from the residual features by SSB. After that, the response maps are produced using the selected seeds and the residual features through DRB. The final response maps and the input residual features are fused to generate the prototype.

where the M^{final} is broadcast to the same size as $\mathcal{F}_{res}$, $\odot$ denotes element-wise multiplication, and $avg(\cdot)$ means averaging feature vector of all the pixels from all inputs.

3.3. Self-Contrastive Learning Module

To further help the DPG to suppress the noise of background, and learn discriminative features without depending on classification information, a self-contrastive learning module (SCL) is designed as shown in Fig. 2. Our motivation is that the prototype generated by the original inputs ($proto$) should be consistent with co-salient prototype generated by inputs where background is erased ($proto_c$), but different from the background prototype generated by inputs where the co-salient objects are erased ($proto_b$). Note that the inputs here are the initial extracted features $\mathcal{F}_{ext}$ from the feature extractor. The co-salient prototype and background prototype can be generated as

$$proto_c = \phi_{\text{DPG}}(\mathcal{F}_{ext} \odot Y^{\downarrow}), \tag{11}$$

$$proto_b = \phi_{\text{DPG}}(\mathcal{F}_{ext} \odot (1 - Y^{\downarrow})), \tag{12}$$

where ϕ_{DPG} is short for the process of DPG, '$\downarrow$' means downscaling the groundtruth Y to the same size as $\mathcal{F}_{ext}$ then broadcasting to the same channel number.

Then, $proto$ and $proto_c$ are treated as a positive pair, while $proto$ and $proto_b$ are treated as a negative pair. A self-contrastive loss is designed to pull together the positive pair and push away the negative pair. First, we define the cosine-style similarity between the prototypes by

$$cos(p_1, p_2) = (1 + \frac{p_1 \cdot p_2}{|p_1| \, |p_2|}) \times 0.5. \tag{13}$$

After that, the self-contrastive loss is defined as

$$cos_c = cos(proto, proto_c), \tag{14}$$

$$cos_b = cos(proto, proto_b), \tag{15}$$

$$\mathcal{L}_{sc} = -log(cos_c + \epsilon) - log(1 - cos_b + \epsilon), \tag{16}$$

where ϵ is a small constant value ensuring non-zero values for $log(\cdot)$ and set as 1×10^{-5}. Our SCL is only applied during training as an auxiliary loss to help the DPG learn more discriminative co-salient features. This part is not used during inference.

3.4. Democratic Feature Enhancement Module

We design a democratic feature enhancement module (DFE) to further strengthen the fused co-salient features from DPG for final prediction. Our DFE is based on the attention mechanism [39]. We observe that conventional attention [39] tends to focus on a limited number of related pixels. Thus, we argue that democracy also matters in this case, and more pixels should be involved in enhancing the fused features. Thus, we try to amplify small positive attention values to involve more pixels for feature enhancement. Negative attention values are not considered here as they usually represent irrelevance. First, we generate the fused features using the guidance of both response maps and the prototype derived from Eq. (9) and Eq. (10) in the DPG as

$$\mathcal{F}_{fused} = \mathcal{F}_{res} \odot M^{final} + \mathcal{F}_{res} \odot proto, \tag{17}$$

where both the M^{final} and $proto$ are broadcast into the same size as $\mathcal{F}_{res}$. Therefore, the fused features by Eq. (17) contain both specific attributes and shared attributes.

The fused features of each input image are enhanced with our DFE individually and independently. As shown in Fig. 4, the corresponding $\mathcal{F}_{fused}$ is input to a 1×1 convolution layer followed by a ReLU activation to obtain $\mathcal{F}_{conv} \in \mathbb{R}^{C \times H \times W}$ first. After that, key, query and
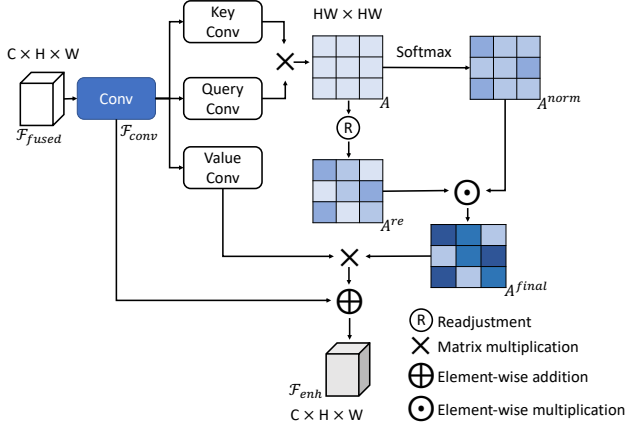
Figure 4. Flow chat of democratic feature enhancement module.

value convolutions are applied and then reshaped to generate $\mathcal{F}_k \in \mathbb{R}^{HW \times C}$, $\mathcal{F}_q \in \mathbb{R}^{HW \times C}$ and $\mathcal{F}_v \in \mathbb{R}^{HW \times C}$. Then, the initial attention map (A) can be computed by

$$A = \mathcal{F}_k \mathcal{F}_q^\top, \qquad (18)$$

where $A \in \mathbb{R}^{HW \times HW}$ and $\top$ means transpose.

Next, a softmax is applied to A to obtain the normalized attention map (A^{norm}). Moreover, the initial attention map A is sorted in a descending order to generate the sorting index matrix (Z). As we adopt the descending order, the small attention values are assigned with large sorting index. Then, we apply the following formula to amplify the small positive attention values,

$$A^{re}_{i,j} = \begin{cases} (Z_{i,j} + 1)^\alpha, & \text{if } A_{i,j} > 0 \\ 1, & \text{else} \end{cases}, \qquad (19)$$

where A^{re} denotes the weights for readjusting the attention, α is a coefficient for determining the degree of amplification, i and j are the spatial index. Then, the final attention map is computed by

$$A^{final} = A^{norm} \odot A^{re}, \qquad (20)$$

and the final enhanced features can be computed by

$$\mathcal{F}_{enh} = \mathcal{F}_{conv} + A^{final} \mathcal{F}_v, \qquad (21)$$

where the result of $A^{final} \mathcal{F}_v$ is first reshaped back into the same size as $\mathcal{F}_{conv}$. Finally, the augmented features $\mathcal{F}_{enh}$ are transmitted into the decoder to predict the corresponding co-saliency maps.

3.5. Objective Function

The objective function for training is a combination of IoU loss [32, 53] and our self-contrastive loss in Eq. (16). The IoU loss can be illustrated as

$$\mathcal{L}_{iou} = 1 - \frac{1}{N} \sum \frac{\hat{Y} \cap Y}{\hat{Y} \cup Y}, \qquad (22)$$

where $\hat{Y}$ denotes for predictions and Y denotes for the groundtruth. Then, the final objective function is

$$\mathcal{L}_{tot} = \mathcal{L}_{iou} + \lambda \mathcal{L}_{sc}, \qquad (23)$$

where λ is to balance IoU loss and self-contrastive loss.

4. Experiment

4.1. Implementation Details

We use Feature Pyramid Network (FPN) [25] with VGG-16 [33] as our backbone. The hyper-parameter α in Eq. (19) is 3 and λ in Eq. (23) is 0.1. Additionally, we use Adam [20] as our optimizer to train our model for 200 epochs. The learning rate is set as 1×10^{-5} for feature extractor and 1×10^{-4} for other parts. The weight decay is set as 1×10^{-4}. In each training episode, we randomly choose one group (16 samples) of relative images. For inference, all samples in each group are input at one time. The inputs are resized into 224×224 for both training and inference. The computational complexity of Eq. (2) is $O((NHW)^2)$ and Eq. (18) is $O((HW)^2)$. The increment of FLOPs is small since the input size is small. The total training time is around 3 hours and the inference time is around 84.4 fps. All experiments are run on one NVIDIA GeForce RTX 2080 Ti.

4.2. Dataset and Evaluation Metrics

Dataset. We use COCO-SEG [40], a subset of COCO dataset [26], which contains 9,213 images from 65 groups for training. We evaluate our method on three popular CoSOD benchmarks: CoCA [54], Cosal2015 [48] and CoSOD3k [10]. CoCA and CoSOD3k are proposed for challenging real-world co-saliency evaluation, containing multiple co-salient objects in some images, large appearance and scale variations, and complex background clutters. Cosal2015 is a widely used large dataset for the evaluation.

Evaluation Metrics. The evaluation metrics include mean absolute error (MAE) [4], maximum F-measure (F_β^{max}) [1], maximum E-measure (E_ϕ^{max}) [8] and S-measure (S_α) [7]. Specifically, the value of MAE is the smaller, the better. While others are the larger, the better.

4.3. Comparison with State-of-The-Art

Compared Methods. We mainly compare with previous state-of-the-art methods trained on common single CoSOD training dataset for fair comparison, including CSMG [50], GCAGC [51], CoEGNet [9], GICD [54], GCoNet [11], and DeepACG [49]. We also list several methods trained on both CoSOD dataset and SOD dataset, such as CADC [52], ICNet [19] and CoADNet [53].

Quantitative Comparison. In Tab. 1, we list the performance comparisons between ours and previous state-of-the-art methods. It can be seen that our method reaches a

Table 1. Comparisons with other state-of-the-art approaches on 3 benchmarks. ↑ means that larger is better and ↓ denotes that smaller is better. 'SOD' denotes training with extra SOD dataset.

| Methods	SOD	CoCA				CoSOD3k				Cosal2015			
		$MAE \downarrow$	$F_\beta^{max} \uparrow$	$E_\xi^{max} \uparrow$	$S_\alpha \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$E_\xi^{max} \uparrow$	$S_\alpha \uparrow$	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$E_\xi^{max} \uparrow$	$S_\alpha \uparrow$
ICNet [19] (NeurIPS20)	✓	0.148	0.506	0.698	0.651	0.097	0.744	0.832	0.780	0.058	0.855	0.900	0.856
CADC [52] (ICCV21)	✓	0.132	0.548	0.744	0.681	0.096	0.759	0.840	0.801	0.064	0.862	0.906	0.866
CoADNet [53] (NeurIPS20)	✓	-	-	-	-	0.070	0.825	-	0.837	0.064	0.875	-	0.861
CSMG [50] (CVPR19)		0.124	0.503	0.734	0.632	0.157	0.645	0.723	0.711	0.130	0.777	0.818	0.774
GCAGC [51] (CVPR20)		0.111	0.523	0.754	0.669	0.100	0.740	0.816	0.785	0.085	0.813	0.866	0.817
CoEGNet [9] (TPAMI21)		0.106	0.493	0.717	0.612	0.092	0.736	0.825	0.762	0.077	0.832	0.882	0.836
GICD [54] (ECCV20)		0.126	0.513	0.715	0.658	0.079	0.770	0.848	0.797	0.071	0.844	0.887	0.844
DeepACG [49] (CVPR21)		0.102	0.552	0.771	0.688	0.089	0.756	0.838	0.792	**0.064**	0.842	0.892	**0.854**
GCoNet [11] (CVPR21)		0.105	0.544	0.760	0.673	0.071	0.777	0.860	0.802	0.068	0.847	0.887	0.845
DCFM (Ours)		**0.085**	**0.598**	**0.783**	**0.710**	**0.067**	**0.805**	**0.874**	**0.810**	0.067	**0.856**	**0.892**	0.838
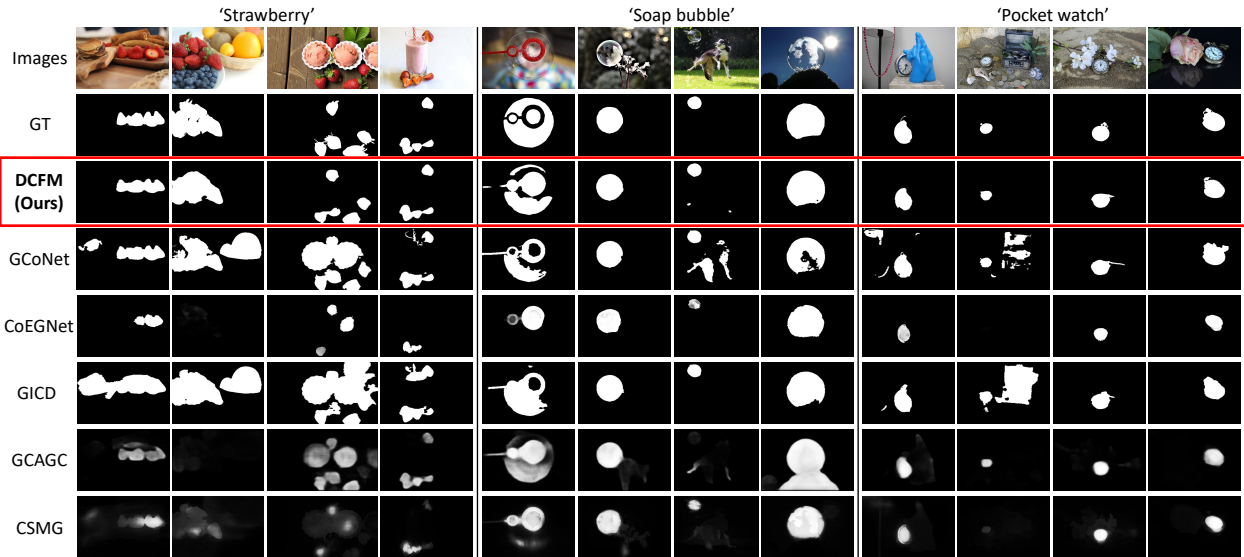
Figure 5. The qualitative comparisons with other state-of-the-art methods. It is evident that our method can predict smoother co-saliency maps with less noise of background, compared with other state-of-the-art methods. More can be found in our supplementary material.

new state-of-the-art performance compared with other approaches under the same settings. Specifically, for the two challenging real-world datasets CoCA and CoSOD3k, *e.g.*, for CoCA, we obtain a gain of 2.0% for MAE, 5.4% for maximum F-measure, 2.3% for maximum E-measure, and 3.7% for S-measure compared with GCoNet [11]. Moreover, our method can even outperform those trained with extra SOD dataset on these two datasets, such as ICNet [19] and CADC [52]. For Cosal2015, our method obtains comparable results with DeepACG [49] and GCoNet [11]. This phenomenon may be caused by the fact that both Deep-ACG [49] and GCoNet [11] use extra classification information to provide structure information, while our method does not rely on any extra information.

Qualitative Comparison. We also report some qualitative comparisons with state-of-the-art methods in Fig. 5. The groups are from CoCA dataset. It can be found that our model can predict more integral and less noisy co-saliency

maps compared with others. Specifically, when there are multiple co-salient objects in one image, like the group 'Strawberry', our model can detect all the target objects, compared with CoEGNet [9] and CSMG [50]. In the group 'Soap bubble', ours are sensitive to appearance variations compared with others. When the background noise level is high, such as the group 'Pocket watch', our predictions contain less noise, compared with GCoNet [11] and GICD [54].

4.4. Ablation Study

We conduct the ablation study of our method on the CoCA dataset by adding one module each time and treating the network with all our modules removed as the baseline. The results are shown in Tab. 2. It can be found that each proposed module contributes a lot. With our DPG, the performance can increase 3.2% for MAE, 5.4% for maximum F-measure, 2.8% for maximum E-measure, and 4.1% for S-measure. Our SCL enables further improvement by 1.0%

Table 2. Ablation study for our proposed modules. 'Base.' denotes baseline. Our overall method obtains the best results.

	Base.	DPG	SCL	DFE	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$E_\xi^{max} \uparrow$	$S_\alpha \uparrow$
1	✓				0.129	0.521	0.735	0.655
2	✓	✓			0.097	0.575	0.763	0.696
3	✓	✓	✓		0.087	0.592	0.775	0.701
4	✓	✓	✓	✓	**0.085**	**0.598**	**0.783**	**0.710**

Table 3. Ablation study for different parts in DPG. 'RB' means the residual block. The overall process obtains the best performance.

	RB	SSB	DRB	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$E_\xi^{max} \uparrow$	$S_\alpha \uparrow$
1				0.129	0.521	0.735	0.655
2	✓			0.124	0.527	0.745	0.659
3	✓	✓		0.126	0.527	0.739	0.657
4	✓	✓	✓	**0.097**	**0.575**	**0.763**	**0.696**

Table 4. Ablation study for different parts in Eq. (16) of SCL. 'cos_c' denotes the case with only positive pair for the loss and 'cos_b' denotes that with only negative pair. DFE is not used.

	cos_c	cos_b	$MAE \downarrow$	$F_\beta^{max} \uparrow$	$E_\xi^{max} \uparrow$	$S_\alpha \uparrow$
1			0.097	0.575	0.763	0.696
2	✓		0.093	0.574	0.764	0.695
3		✓	0.095	0.583	0.773	0.697
4	✓	✓	**0.087**	**0.592**	**0.775**	**0.701**

Table 5. Ablation study for readjustment in DFE. 'w/o DFE' denotes not using DFE, 'w/o RA' denotes using DFE without readjustment and 'w/ RA' denotes using DFE with readjustment.

		$MAE \downarrow$	$F_\beta^{max} \uparrow$	$E_\xi^{max} \uparrow$	$S_\alpha \uparrow$
1	w/o DFE	0.087	0.592	0.775	0.701
2	w/o RA	0.100	0.567	0.769	0.691
3	w/ RA	**0.085**	**0.598**	**0.783**	**0.710**

for MAE, 1.7% for maximum F-measure, 1.2% for maximum E-measure, and 0.5% for S-measure, respectively. Besides, our model with DFE reach 0.085 for MAE, 0.598 for maximum F-measure, 0.783 for maximum E-measure, and 0.710 for S-measure. The new state-of-the-art performance is obtained when all the modules are included.

Impact of Democratic Prototype Generation Module. The evaluation of each block of our DPG is listed in Tab. 3. The experiment is conducted by adding one block at a time. Compared with the baseline (row 1), each part of DPG devotes to the final results. Specifically, if we only use RB and SSB, where we take the mean of seeds as the prototype, the results are even lower than the case without SSB, comparing row 2 and 3. On the other hand, with DRB, comparing row 3 and 4, the results will be increased by 2.9% for MAE, 4.8% for maximum F-measure, 2.4% for maximum E-measure, and 3.9% for S-measure. This phenomenon can verify that democracy does matter. More co-salient pixels should be enrolled for the comprehensive prototype.

Impact of Self-Contrastive Learning Module. We also evaluate two main parts in our self-contrastive loss as listed in Tab. 4. We conduct this experiment by removing one part each time. It can be seen that with only positive pair cos_c, by comparing row 1 and 2, we can get comparable results.

With only negative pair cos_b, the performance is clearly improved, by comparing row 1 and 3. This phenomenon proves that the negative pair is important for removing background noise. Nevertheless, by comparing row 3 and 4, the contrastive learning with both positive and negative pairs promotes balanced training for higher results. More analysis can be found in our supplementary material.

Impact of Democratic Feature Enhancement Module. We also experiment on the readjustment of attention values in Tab. 5. When the readjustment is removed but using conventional attention in our DFE, the performance is even worse than the case without our DFE, as shown in row 1 and 2. Thus, democracy does matter in this module as well. Conventional attention mechanism focusing on limited pixels cannot provide sufficient information for the decoder while more related pixels should be involved.

5. Conclusion

In this paper, we have proposed a new method for CoSOD without using the SOD dataset and classification information. We design a democratic prototype generation module (DPG) to build democratic response maps first so as to generate a comprehensive prototype as guidance for further prediction. Moreover, to help suppress noisy background information in the prototype, we design a self-contrastive learning module (SCL), where both positive and negative pairs are generated from the image itself without relying on classification information. Besides, we also design a democratic feature enhancement module (DFE) to strengthen co-salient features from DPG for final prediction. Both our DPG and DFE show that democracy does matter. More related pixels should be involved for mining comprehensive features for CoSOD.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009. 6

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Int. Conf. Mach. Learn.*, 2020. 2

[3] Zuyao Chen, Qianqian Xu, Runmin Cong, and Qingming Huang. Global context-aware progressive aggregation network for salient object detection. In *AAAI*, 2020. 1

[4] Ming-Ming Cheng, Jonathan Warrell, Wen-Yan Lin, Shuai Zheng, Vibhav Vineet, and Nigel Crook. Efficient salient region detection with soft image abstraction. In *Int. Conf. Comput. Vis.*, 2013. 6

[5] Runmin Cong, Jianjun Lei, Huazhu Fu, Ming-Ming Cheng, Weisi Lin, and Qingming Huang. Review of visual saliency detection with comprehensive information. *IEEE Trans. Circuit Syst. Video Technol.*, 29(10):2941–2959, 2018. 2

[6] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Int. Conf. Comput. Vis.*, 2021. 2

[7] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: a new way to evaluate foreground maps. In *Int. Conf. Comput. Vis.*, 2017. 6

[8] 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*, 2019. 6

[9] Deng-Ping Fan, Tengpeng Li, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Ming-Ming Cheng, Huazhu Fu, and Jianbing Shen. Re-thinking co-salient object detection. *arXiv preprint*, 2020. 2, 6, 7

[10] Deng-Ping Fan, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Huazhu Fu, and Ming-Ming Cheng. Taking a deeper look at co-salient object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 6

[11] Qi Fan, Deng-Ping Fan, Huazhu Fu, Chi-Keung Tang, Ling Shao, and Yu-Wing Tai. Group collaborative learning for co-salient object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 2, 6, 7

[12] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 3

[13] Guangshuai Gao, Wenting Zhao, Qingjie Liu, and Yunhong Wang. Co-saliency detection with co-attention fully convolutional network. *IEEE Trans. Circuit Syst. Video Technol.*, 31(3):877–889, 2020. 3

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[15] Rongyao Hu, Zhenyun Deng, and Xiaofeng Zhu. Multi-scale graph fusion for co-saliency detection. In *AAAI*, 2021. 2

[16] Bo Jiang, Xingyue Jiang, Jin Tang, and Bin Luo. Co-saliency detection via a general optimization model and adaptive graph learning. *IEEE Trans. Multimedia*, 23:3193–3202, 2021. 2

[17] Bo Jiang, Xingyue Jiang, Jin Tang, Bin Luo, and Shilei Huang. Multiple graph convolutional networks for co-saliency detection. In *Int. Conf. Multimedia and Expo*, 2019. 2

[18] 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 Int. Conf. Multimedia*, 2019. 2

[19] Wen-Da Jin, Jun Xu, Ming-Ming Cheng, Yi Zhang, and Wei Guo. Icnet: intra-saliency correlation network for co-saliency detection. In *Adv. Neural Inform. Process. Syst.*, 2020. 1, 2, 6, 7

[20] Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *Int. Conf. Learn. Represent.*, 2014. 6

[21] Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1

[22] Bo Li, Zhengxing Sun, Lv Tang, Yunhan Sun, and Jinlong Shi. Detecting robust co-saliency with recurrent co-attention neural network. In *IJCAI*, 2019. 3

[23] Hui Li, Jimin Xiao, Mingjie Sun, Eng Gee Lim, and Yao Zhao. Transformer based language-person search with multiple region slicing. *IEEE Trans. Circuit Syst. Video Technol.*, 2021. 3

[24] Jia Li, Jinming Su, Changqun Xia, Mingcan Ma, and Yonghong Tian. Salient object detection with purificatory mechanism and structural similarity loss. *IEEE Trans. Image Process.*, 30:6855–6868, 2021. 1

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 6

[26] 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 *Eur. Conf. Comput. Vis.*, 2014. 6

[27] Yun Liu, Xin-Yu Zhang, Jia-Wang Bian, Le Zhang, and Ming-Ming Cheng. Samnet: stereoscopically attentive multi-scale network for lightweight salient object detection. *IEEE Trans. Image Process.*, 30:3804–3814, 2021. 1

[28] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[29] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Int. Conf. Comput. Vis.*, 2019. 3

[30] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1

[31] Yongri Piao, Jian Wang, Miao Zhang, and Huchuan Lu. Mfnet: multi-filter directive network for weakly supervised salient object detection. In *Int. Conf. Comput. Vis.*, 2021. 1

[32] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: boundary-aware salient object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 6

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Int. Conf. Learn. Represent.*, 2014. 6

[34] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[35] Mingjie Sun, Jimin Xiao, Eng Gee Lim, Si Liu, and John Y Goulermas. Discriminative triad matching and reconstruction for weakly referring expression grounding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(11):4189–4195, 2021. 3

[36] Lv Tang, Bo Li, Yijie Zhong, Shouhong Ding, and Mofei Song. Disentangled high quality salient object detection. In *Int. Conf. Comput. Vis.*, 2021. 1

[37] Janine Thoma, Danda Pani Paudel, and Luc V Gool. Soft contrastive learning for visual localization. In *Adv. Neural Inform. Process. Syst.*, 2020. 2

[38] Chung-Chi Tsai, Kuang-Jui Hsu, Yen-Yu Lin, Xiaoning Qian, and Yung-Yu Chuang. Deep co-saliency detection via stacked autoencoder-enabled fusion and self-trained cnns. *IEEE Trans. Multimedia*, 22(4):1016–1031, 2019. 2

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, 2017. 1, 2, 3, 5

[40] Chong Wang, Zheng-Jun Zha, Dong Liu, and Hongtao Xie. Robust deep co-saliency detection with group semantic. In *AAAI*, 2019. 6

[41] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In *Eur. Conf. Comput. Vis.*, 2020. 3

[42] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 3

[43] 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 Trans. Image Process.*, 28(10):5052–5063, 2019. 2

[44] Siyue Yu, Jimin Xiao, Bingfeng Zhang, Eng Gee Lim, and Yao Zhao. Fast pixel-matching for video object segmentation. *Sign. Process. Image Communication*, 98:116373, 2021. 3

[45] Siyue Yu, Bingfeng Zhang, Jimin Xiao, and Eng Gee Lim. Structure-consistent weakly supervised salient object detection with local saliency coherence. In *AAAI*, 2021. 1

[46] Bingfeng Zhang, Jimin Xiao, Jianbo Jiao, Yunchao Wei, and Yao Zhao. Affinity attention graph neural network for weakly supervised semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021. 3

[47] Dingwen Zhang, Huazhu Fu, Junwei Han, Ali Borji, and Xuelong Li. A review of co-saliency detection algorithms: fundamentals, applications, and challenges. *ACM Trans.Intell. Syst. and Tech.*, 9(4):1–31, 2018. 2

[48] Dingwen Zhang, Junwei Han, Chao Li, Jingdong Wang, and Xuelong Li. Detection of co-salient objects by looking deep and wide. *Int. J. Comput. Vis.*, 120(2):215–232, 2016. 6

[49] Kaihua Zhang, Mingliang Dong, Bo Liu, Xiao-Tong Yuan, and Qingshan Liu. Deepacg: co-saliency detection via semantic-aware contrast gromov-wasserstein distance. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 2, 6, 7

[50] Kaihua Zhang, Tengpeng Li, Bo Liu, and Qingshan Liu. Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 6, 7

[51] Kaihua Zhang, Tengpeng Li, Shiwen Shen, Bo Liu, Jin Chen, and Qingshan Liu. Adaptive graph convolutional network with attention graph clustering for co-saliency detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 6, 7

[52] Ni Zhang, Junwei Han, Nian Liu, and Ling Shao. Summarize and search: learning consensus-aware dynamic convolution for co-saliency detection. In *Int. Conf. Comput. Vis.*, 2021. 1, 2, 6, 7

[53] Qijian Zhang, Runmin Cong, Junhui Hou, Chongyi Li, and Yao Zhao. Coadnet: collaborative aggregation-and-distribution networks for co-salient object detection. In *Adv. Neural Inform. Process. Syst.*, 2020. 1, 2, 6, 7

[54] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In *Eur. Conf. Comput. Vis.*, 2020. 2, 6, 7

[55] Ziqi Zhou, Zheng Wang, Huchuan Lu, Song Wang, and Meijun Sun. Multi-type self-attention guided degraded saliency detection. In *AAAI*, 2020. 3