This CVPR 2021 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Group Collaborative Learning for Co-Salient Object Detection

Qi Fan[1,3,*] Deng-Ping Fan[2,†] Huazhu Fu[2] Chi-Keung Tang[1] Ling Shao[2] Yu-Wing Tai[1,3]
[1] HKUST [2] Inception Institute of Artificial Intelligence [3] Kuaishou Technology
fanqics@gmail.com, dengpfan@gmail.com, hzfu@ieee.org,
cktang@cs.ust.hk, ling.shao@inceptioniai.org, yuwing@gmail.com

Abstract

*We present a novel group collaborative learning framework (**GCoNet**) capable of detecting co-salient objects in real time (16ms), by simultaneously mining consensus representations at group level based on the two necessary criteria: **1) intra-group compactness** to better formulate the consistency among co-salient objects by capturing their inherent shared attributes using our novel group affinity module; **2) inter-group separability** to effectively suppress the influence of noisy objects on the output by introducing our new group collaborating module conditioning the inconsistent consensus. To learn a better embedding space without extra computational overhead, we explicitly employ auxiliary classification supervision. Extensive experiments on three challenging benchmarks,* i.e., *CoCA, CoSOD3k, and Cosal2015, demonstrate that our simple GCoNet outperforms 10 cutting-edge models and achieves the new state-of-the-art. We demonstrate this paper's new technical contributions on a number of important downstream computer vision applications including content aware co-segmentation, co-localization based automatic thumbnails, etc. Code has been made publicly available: https://github.com/fanq15/GCoNet.*

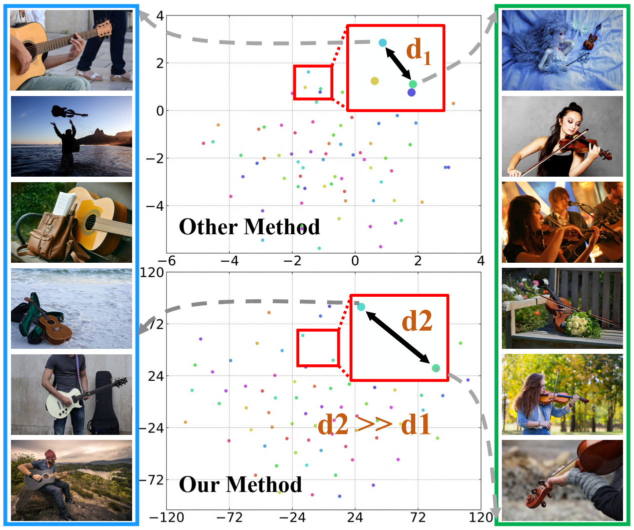

Figure 1. **t-SNE [18] visualization of consensuses**, where each point represents one consensus of an image group. Highlighted here are two similar but different groups (*guitar* & *violin*) to illustrate the effectiveness of GCoNet. The consensus strategy in traditional CoSOD model (CoEGNet [8]) tends to cluster consensuses together even they belong to different groups, resulting in ambiguous co-saliency detection. In contrast, our consensus strategy with effective inter-group constraint enables higher diversity with a very large group variance ($d_2 \gg d_1$) and thus better inter-group separability.

1. Introduction

Co-salient object detection (CoSOD) targets at detecting common salient objects sharing the same attributes given a group of relevant images. CoSOD is more challenging than the standard salient object detection (SOD) task [1, 2, 3] and RGB-D SOD [4, 5, 6, 7], because CoSOD needs to distinguish co-occurring objects across multiple images [8] in presence of other objects. That is, both intra-class compactness and inter-class separability should be simultaneously maximized. With this favorable feature CoSOD is thus often employed as a pre-processing step for various computer vision tasks, such as image retrieval [9], image quality assessment [10], collection-based crops [11], co-segmentation [12, 13], semantic segmentation [14], image surveillance [15], video analysis [16], video co-localization [17], *etc*.

Previous works attempt to leverage the consistency among relevant images to facilitate CoSOD *within an image group* by exploring different shared cues [19, 20, 21] or semantic connections [22, 23, 24]. Some of them [25, 26] use predicted saliency maps by computing various inter-image cues to discover co-salient objects. Other works [8, 27] exploit a unified network to jointly optimize co-saliency information and saliency maps.

[*] This work was done when Qi was an intern at Kuaishou Technology.
[†] Corresponding author: Deng-Ping Fan *(dengpfan@gmail.com)*.

Despite their promising results, most current models only extract their CoSOD representations in an individual group, which introduces a number of limitations. First, images from the same group contain similar foregrounds (*i.e.*, co-salient objects) only provide positive relations while lacking the negative relations between different objects. Training the model only using positive pairs may lead to overfitting and result in ambiguous results for outlier images. Moreover, the number of images in a group is typically limited (20 to 40 images for most CoSOD datasets), so using a single group cannot provide enough information for learning a discriminative representation. Finally, individual groups also fall short in offering high-level semantic information, which is necessary for distinguishing noisy objects during inference in complex real-world scenarios.

To address the above issues, we propose a novel group collaborative learning framework (**GCoNet**) to mine the semantic correlation between *different image groups*. The proposed GCoNet consists of three important components: group affinity module (GAM), group collaborating module (GCM) and auxiliary classification module (ACM), which simultaneously learn the **intra-group compactness** and **inter-group separability**. The GAM makes the network learn the consensus feature within the same image group, while the GCM discriminates target attributes between different groups, thus enabling the model to be trained on the existing large-scale SOD datasets.[1] We further improve the feature representation at a global semantic level through our ACM on each image to learn a better embedding space. In summary, our contributions are:

- We introduce a novel group collaborative learning strategy to address the CoSOD problem, and validate its effectiveness with extensive ablation studies.
- We design a novel unified Group Collaborative Learning Network (**GCoNet**) for CoSOD by simultaneously considering intra-group compactness and inter-group separability to mine the consensus representation.
- Our group affinity module (GAM) and group collaborating module (GCM) collaborate with each other to achieve better intra- and inter-group collaborative learning. The auxiliary classification module (ACM) further promotes learning at a global semantic level.
- Extensive experiments on three challenging CoSOD benchmarks, *i.e.*, CoCA, CoSOD3k, and Cosal2015, show that our GCoNet achieves the new state-of-the-art. Furthermore, we present two downstream applications based on our technical contributions, *i.e.*, cosegmentation and co-localization.

2. Related Work

The traditional salient object detection task [28, 29, 30, 31, 32] targets at directly segmenting salient object in each image separately, while CoSOD aims to segment the common salient objects across several relevant images. Previous works mainly exploit inter-image cues to detect co-salient objects. Early CoSOD methods explore the inter-image correspondence between image-pairs [19, 33] or a group of relevant images [34] based on shallow handcrafted descriptors [17, 35]. They employ different approaches to mine the inter-image relationships using constraints or heuristic characteristics. Several studies attempt to capture the inter-image constraints by employing an efficient manifold ranking scheme [36] to obtain guided saliency maps, or using a global association constraint with clustering [20], or translational alignment [11]. Other works attempt to formulate the semantic attributes shared among images in a group from the high-level features in the heuristic characteristics, using multiple saliency cues and self-adaptive weights [21], regional histograms and constrasts [37], metric learning by optimizing a new objective function [23], or pairwise similarity ranking and linear programming [38].

Recently deep-based models simultaneously explore the intra- and inter-image consistency in a supervised manner with different approaches, such as graph convolution networks (GCN) [39, 40, 41], self-learning methods [22, 42], inter-image co-attention with PCA projection [8] or recurrent units [43], correlation techniques [44], quality measurement [45], or co-clustering [46]. Some methods exploit multi-task learning to simultaneously optimize the co-saliency detection and co-segmentation [47] or co-peak search [13]. Other works explore hierachical features from multi-scale [48], multi-stage [49], or multi-layer [50] features. Another notable research line is to explore groupwise semantic representation (consensus) which is used to detect co-salient regions for each image. There are different methods to capture the discriminative semantic representation, such as group attentional semantic aggregation [51], gradient feedback [27], co-category association [52], united fully convolutional network [53, 54], or integrated multilayer graph [55]. Methods are proposed to solve the CoSOD problem in a semi-supervised [56] or unsupervised manner [24, 57, 58, 59], and studies [60, 61] are availalbe on co-saliency detection from a single image.

Previous works have focused on intra-group (intra- and inter-image) cues for capturing common attributes of co-salient objects. The inter-group information has received less attention, although CODW [62] focuses on *visually similar* neighbor. Recently Zhang *et al.* [27] utilized a jigsaw training to *implicitly* exploit other images to facilitate group training. But their model still targets intra-group learning. Our method differs from existing models in the exploration of inter-group relations for discriminating fea-

[1] Note that the existing CoSOD datasets altogether contain about 6k images, while there are more than 12 SOD datasets, containing about 60k images. It may partially alleviate the insufficient training data issue in co-salient object detection.

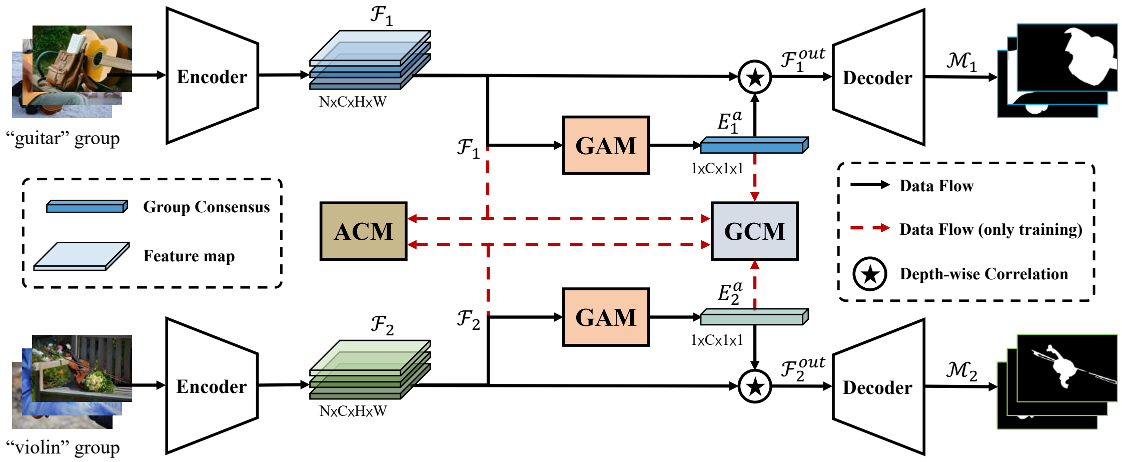

Figure 2. **Pipeline of the proposed Group Collaborative Learning Network (GCoNet).** Images in two groups are first processed by a weight-shared encoder. Then we employ the group affinity module (GAM, see Figure 3 for more details) to conduct intra-group collaborative learning for each group to generate a consensus, which is collaborated with the original feature maps to segment co-salient objects using the decoder. In addition, the original feature maps and consensuses of both groups are fed to the group collaborating module (GCM, see Figure 4) to conduct the inter-group collaborative learning. Moreover an auxiliary classification module (ACM) is applied to obtain the high-level semantic representation. The GCM and ACM are only used for training and are removed at inference.

ture learning at a group level *explicitly and semantically*.

3. Group Collaborative Learning Network

3.1. Architecture Overview

Given a group of N relevant images $\{I_1, I_2, ..., I_n\}$ containing common salient objects of a certain class, CoSOD aims to detect them simultaneously and output the co-saliency maps. Unlike existing CoSOD methods which only depend on the information within the image group, we propose a novel group collaborative learning network (GCoNet) to mine the consensus representations at both intra- and inter-group level.

Figure 2 illustrates the flowchart of our GCoNet. First, an encoder network is used to extract feature maps $\mathcal{F}_1 = \{F_{1,n}\}_{n=1}^N, \mathcal{F}_2 = \{F_{2,n}\}_{n=1}^N \in \mathbb{R}^{N\times C\times H\times W}$ for two image groups, where C is the channel number and $H \times W$ is the spatial size. Then, a group affinity module (GAM) is used to combine all single-image features to distill the consensus feature $E_1^a, E_2^a \in \mathbb{R}^{1\times C\times 1\times 1}$ ($C = 512$ in our experiments) from $\mathcal{F}_1, \mathcal{F}_2$, representing the common attributes of the co-salient objects for each group. Simultaneously, a group collaborating module (GCM) is applied to enhance the image representation for discriminating the target attributes between different image groups. Finally, we further improve the high-level semantic representation of images using an auxiliary classification module (ACM) to learn a better embedding space. The resulting collaborative features are then fed to a decoder network to produce the co-saliency maps $\mathcal{M}_1, \mathcal{M}_2$.

3.2. Group Affinity Module

Intuitively, common objects from the same class always share some similarity in appearance and have high similarity in features, which have been widely employed in many tasks. Inspired by self-supervised video tracking methods [63, 64, 65, 66], which propagate the segmentation masks of target objects based on the pixel-wise correspondences between two adjacent frames, we extend this idea to the CoSOD task by computing the global affinity among all images in a group.

For any two image features $\{F_{1,n}, F_{1,m}\} \in \mathcal{F}_1$ [2] and without losing generality we drop the group subscript, we can use the inner product to compute their pixel-wise correlations:

$$S_{(n,m)} = \theta(F_n)^T \phi(F_m), \qquad (1)$$

where θ, ϕ are linear embedding functions ($3 \times 3 \times 512$ convolutional layer). The affinity map $S_{(n,m)} \in \mathbb{R}^{HW\times HW}$ efficiently captures the commonality of co-salient objects in the image pair (n, m). Then we can generate F_n's affinity map $A_{n\leftarrow m} \in \mathbb{R}^{HW\times 1}$ by finding the maxima for each of F_n's pixel conditioned on F_m which alleviates the influence of noisy correlation values in the map.

Similarly, we can extend the local affinity of two images to the global affinity of all images in the group. Specifically, we compute the affinity map $S_{\mathcal{F}} \in \mathbb{R}^{NHW\times NHW}$ between all image features $\mathcal{F}$ using Eq. 1. Then, we find the maxima for each image $A'_{\mathcal{F}} \in \mathbb{R}^{NHW\times N}$ from $S_{\mathcal{F}}$, and average all the maxima of N images to generate the global affinity attention map $A_{\mathcal{F}} \in \mathbb{R}^{NHW\times 1}$. In this way, the affinity atten-

[2] All analyses in section 3.2 on $\mathcal{F}_1$ can be applied to $\mathcal{F}_2$. We omit the group subscript for notation simplicity, *i.e.*, we use F_n to represent $F_{1,n}$.

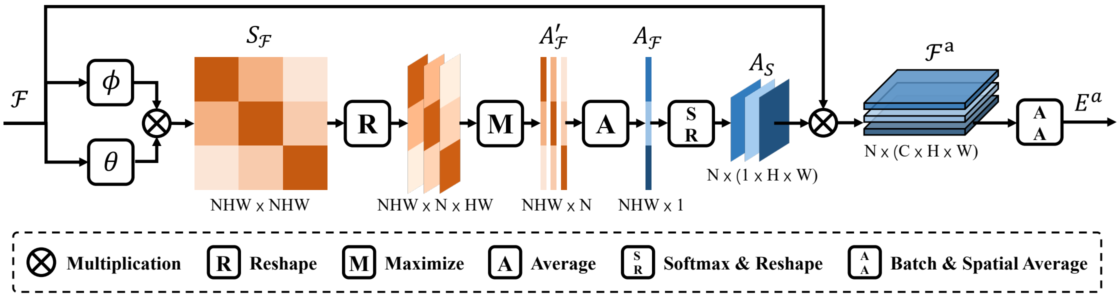

Figure 3. **Group Affinity Module.** We first exploit the affinity attention to generate the attention maps for the input features by collaborating all images in group. Subsequently, the maps are multiplied with the input features to generate the consensus for the group. Then the obtained consensus is used to coordinate the original feature maps and is also fed to the GCM for inter-group collaborative learning.

tion map is globally optimized on all images thus alleviating the influence of occasional co-occurring bias. Then, we use a softmax operation to normalize $A_{\mathcal{F}}$ and reshape it to generate the attention map $A_S \in \mathbb{R}^{N\times(1\times H\times W)}$. We multiply A_S with the original feature $\mathcal{F}$ to produce the attention feature maps $\mathcal{F}^a \in \mathbb{R}^{N\times C\times H\times W}$. Finally, all the attention feature maps $\mathcal{F}^a$ for the whole group are used to produce the attention consensus E^a by average pooling along both the batch and spatial dimensions, as shown in Figure 3.

The global affinity module focuses on capturing the commonality among co-salient objects within the same group and therefore improves the intra-group compactness of the consensus representation. Such *intra-group compactness* alleviates the disturbance of co-occurring noise and enables the model to concentrate on the co-salient regions. This allows the shared attributes of co-salient objects to be better captured and therefore results in better consensus representation. The obtained attention consensus E^a is combined with the original feature maps $\mathcal{F}$ through depth-wise correlation [67, 68] to achieve efficient information association. The resulting feature maps $\mathcal{F}^{out}$ are fed to the decoder to predict co-saliency maps $\mathcal{M}_n$ for each image. The loss function is:

$$\mathcal{L}_{\text{sal}} = \frac{1}{N}\sum_{n}^{N}\mathcal{L}_{\text{siou}}(\mathcal{M}_n, \mathcal{G}_n), \tag{2}$$

where $\mathcal{L}_{\text{siou}}$ is the soft IoU loss [28, 69] and $\mathcal{G}_n$ denotes the ground-truth label for each image in the group.

3.3. Group collaborating module (GCM)

Most CoSOD methods tend to focus on the intra-group compactness of the consensus, but the *inter-group separability* is equally crucial for distinguishing distracting objects, especially when processing complex images with more than one salient objects. To enhance the discriminative representations between different groups, we propose a simple but effective module, *i.e.*, the GCM, by learning to encode the inter-group separability.

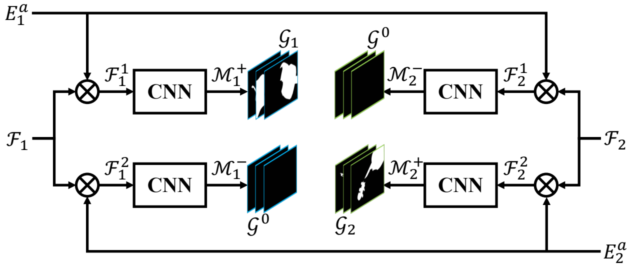

Figure 4. **Group collaborating module.** The original feature maps and consensuses of both groups are fed to the GCM. The predicted output conditioned on the consistent feature and consensus (from the same group) is supervised with the available ground-truth labels. Otherwise, it is supervised by the all-zero maps.

Given two image groups with the corresponding features $\{\mathcal{F}_1, \mathcal{F}_2\}$ and attention consensus $\{E_1^a, E_2^a\}$ obtained from the GAM, we apply an intra- and inter-group cross-multiplication. Specifically, the intra-group multiplication deals with the features and their consensus: $\mathcal{F}_1^1 = \mathcal{F}_1 \cdot E_1^a$ and $\mathcal{F}_2^2 = \mathcal{F}_2 \cdot E_2^a$ for the intra-group collaboration, while the inter-group multiplication acts on the features and consensus of different groups, *i.e.*, $\mathcal{F}_1^2 = \mathcal{F}_1 \cdot E_2^a$ and $\mathcal{F}_2^1 = \mathcal{F}_2 \cdot E_1^a$, to express the inter-group interaction. The intra-group representation $\mathcal{F}^+ = \{\mathcal{F}_1^1, \mathcal{F}_2^2\}$ is exploited to predict the co-saliency maps, and the inter-group representation $\mathcal{F}^- = \{\mathcal{F}_1^2, \mathcal{F}_2^1\}$ is employed to provide a consensus with group separability. Specifically, we feed $\{\mathcal{F}^+, \mathcal{F}^-\}$ to a small convolutional network with an upsampling layer and produce the saliency map $\{\mathcal{M}^+, \mathcal{M}^-\}$[3] with different supervision signals: we use ground-truth labels to supervise $\mathcal{F}^+$, while all-zero maps are used for $\mathcal{F}^-$. The loss function is:

$$\mathcal{L}_{\text{ctm}} = \frac{1}{N}\sum_{n}^{N}\mathcal{L}_{\text{FL}}(<\mathcal{M}_n^+, \mathcal{M}_n^->, <\mathcal{G}_n, \mathcal{G}_n^0>), \tag{3}$$

where $\mathcal{L}_{\text{FL}}$ is the focal loss [70], $\mathcal{G}_n$ is the ground-truth, $\mathcal{G}_n^0$ is the all-zero map and $< \cdot >$ denotes the concatenation operation.

Our GCM thus encourages the consensus to distinguish different groups with high inter-group separability to iden-

[3] $\mathcal{M}^+ = \{\mathcal{M}_1^+, \mathcal{M}_2^+\}$ and $\mathcal{M}^- = \{\mathcal{M}_1^-, \mathcal{M}_2^-\}$.

Table 1. **Quantitative ablation studies** of our GCoNet on the effectiveness of the GAM (group affinity module), GCM (group collaborating module), ACM (auxiliary classification module) and their combinations.

	Modules			CoCA [27]				CoSOD3k [8]				Cosal2015 [62]			
ID	GAM	GCM	ACM	$E_\phi^{\max} \uparrow$	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$\epsilon \downarrow$	$E_\phi^{\max} \uparrow$	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$\epsilon \downarrow$	$E_\phi^{\max} \uparrow$	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$\epsilon \downarrow$
1				0.618	0.591	0.419	0.190	0.811	0.764	0.721	0.108	0.862	0.818	0.800	0.087
2	✓			0.663	0.605	0.442	0.160	0.823	0.772	0.736	0.099	0.873	0.825	0.815	0.079
3		✓		0.666	0.616	0.452	0.156	0.839	0.788	0.748	0.087	0.877	0.834	0.823	0.074
4			✓	0.651	0.606	0.442	0.167	0.829	0.779	0.737	0.094	0.875	0.832	0.820	0.076
5	✓	✓		0.719	0.650	0.504	0.126	0.850	0.798	0.766	0.078	0.884	0.842	0.837	0.070
	✓	✓	✓	**0.760**	**0.673**	**0.544**	**0.105**	**0.860**	**0.802**	**0.777**	**0.071**	**0.888**	**0.845**	**0.847**	**0.068**

tify distractors in complex environment. Another advantage is that this module enables the model to be trained on the existing SOD datasets, whose images typically contain only one dominating object. We can discard this module during inference without introducing additional computational overhead.

3.4. Auxiliary Classification Module (ACM)

To obtain more discriminative features for consensus, we also introduce an ACM to facilitate high-level semantic representation learning. Specifically, we add a classification predictor with a global average pooling layer and one fully connected layer to the backbone to classify F_n to the corresponding class $\mathcal{Y}_n$. In the Euclidean feature space, the classification supervision can separate classes by introducing a large margin, and cluster samples belonging to the same class. Therefore, it enables the model to generate more representative features and benefits the consensus learning for intra-group compactness and inter-group separability. The loss function is:

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{ce}}(\mathcal{Y}_n, \hat{\mathcal{Y}}_n), \qquad (4)$$

where $\mathcal{L}_{\text{ce}}$ is the cross-entropy loss and $\hat{\mathcal{Y}}_n$ is the ground-truth class label.

3.5. End-to-end Training

During training, the GAM, GCM, and ACM are jointly trained with the backbone in an end-to-end manner. The whole framework is optimized by integrating all the aforementioned loss functions:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{sal}} + \lambda_2 \mathcal{L}_{\text{ctm}} + \lambda_3 \mathcal{L}_{\text{cls}}, \qquad (5)$$

where λ_1, λ_2, and λ_3 are hyperparameter weights to balance the loss functions.

4. Experiments

4.1. Implementation Details

We use VGG-16 [71] with Feature Pyramid Network (FPN) [72] as our backbone. For fair comparison, we follow GICD [27] and use the DUTS [73] dataset as our training set. The group labels derived from GICD [27] are used to group the images during training. In each training episode, we randomly pick two different groups with 16 samples[4] in each group to train the network. The images are all resized to 224x224 for training and testing, and the output saliency maps are resized to the original size for evaluation. The network is trained over 50 epochs in total with the Adam optimizer. The initial learning rate is set to $10e-4$, $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The whole training takes around four hours and the inference speed on the image pair groups[5] is 16 ms. The platform for training and inference is equipped with 56 Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz and a Nvidia GeForce GTX 1080Ti.

4.2. Evaluation Datasets and Metrics

We employ three challenging datasets for evaluation: CoCA [27], CoSOD3k [74], and Cosal2015 [62]. The last is a large dataset widely used in the evaluation of CoSOD methods. The first two were recently proposed for challenging real-world co-saliency evaluation, with the images usually containing multiple common and non-common objects against a complex background. Following the advice of recent large-scale benchmark work [74], we do not use iCoseg [75] and MSRC [76] for evaluation, because they usually provide only one salient object in an image and are not very suitable for evaluating CoSOD models. We use maximum E-measure $E_\phi^{\max}$ [77], S-measure S_α [78], maximum F-measure $F_\beta^{\max}$ [79], and mean absolute error (MAE) ϵ [80] to evaluate methods in our experiments.[6]

4.3. Ablation Studies

In this section, we study the effectiveness of each component in our approach (Table 1) and investigate how they contribute to a good consensus feature.

Effectiveness of GAM. The global co-attention module is a fundamental component of our model, which is designed to capture the common attributes of co-salient objects in an image group for better *intra-group compactness*. Compared to the baseline model with only the vanilla consensus extracted by an average pooling operation, GAM improves the performance on all metrics and datasets. To get a deeper un-

[4] Due to limited computing resource. The larger the better.

[5] CoSOD task works for image groups. Therefore we use the basic image pair group to evaluate the speed rather than the single image.

[6] Evaluation toolbox: https://github.com/DengPingFan/CoSODToolbox.

Table 2. **Quantitative comparison results** between our GCoNet and other methods. "↑" ("↓") means that the higher (lower) is better. Co = CoSOD models, Sin = Single-SOD models. The symbol $*$ denotes traditional CoSOD algorithms.

			CoCA [27]				CoSOD3k [8]				Cosal2015 [62]			
Method	Pub. & Year	Type	$E_\phi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$	$E_\phi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$	$E_\phi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$
CBCD* [20]	TIP 2013	Co	0.641	0.523	0.313	0.180	0.637	0.528	0.466	0.228	0.656	0.544	0.532	0.233
GWD [54]	IJCAI 2017	Co	0.701	0.602	0.408	0.166	0.777	0.716	0.649	0.147	0.802	0.744	0.706	0.148
RCAN [43]	IJCAI 2019	Co	0.702	0.616	0.422	0.160	0.808	0.744	0.688	0.130	0.842	0.779	0.764	0.126
CSMG [48]	CVPR 2019	Co	0.733	0.627	0.499	0.114	0.804	0.711	0.709	0.157	0.842	0.774	0.784	0.130
BASNet [28]	CVPR 2019	Sin	0.644	0.592	0.408	0.195	0.804	0.771	0.720	0.114	0.849	0.822	0.791	0.096
PoolNet [29]	CVPR 2019	Sin	0.640	0.602	0.404	0.177	0.799	0.771	0.709	0.113	0.848	0.823	0.785	0.094
EGNet [30]	ICCV 2019	Sin	0.648	0.603	0.404	0.178	0.793	0.762	0.702	0.119	0.843	0.818	0.786	0.099
SCRN [81]	ICCV 2019	Sin	0.642	0.612	0.413	0.164	0.805	0.771	0.716	0.113	0.850	0.817	0.783	0.098
GICD [27]	ECCV 2020	Co	0.715	0.658	0.513	0.126	0.848	0.797	0.770	0.079	**0.887**	0.844	0.844	0.071
CoEGNet [8]	TPAMI 2021	Co	0.717	0.612	0.493	0.106	0.825	0.762	0.736	0.092	0.882	0.836	0.832	0.077
GCoNet (Ours)	CVPR 2021	Co	**0.760**	**0.673**	**0.544**	**0.105**	**0.860**	**0.802**	**0.777**	**0.071**	**0.887**	**0.845**	**0.847**	**0.068**

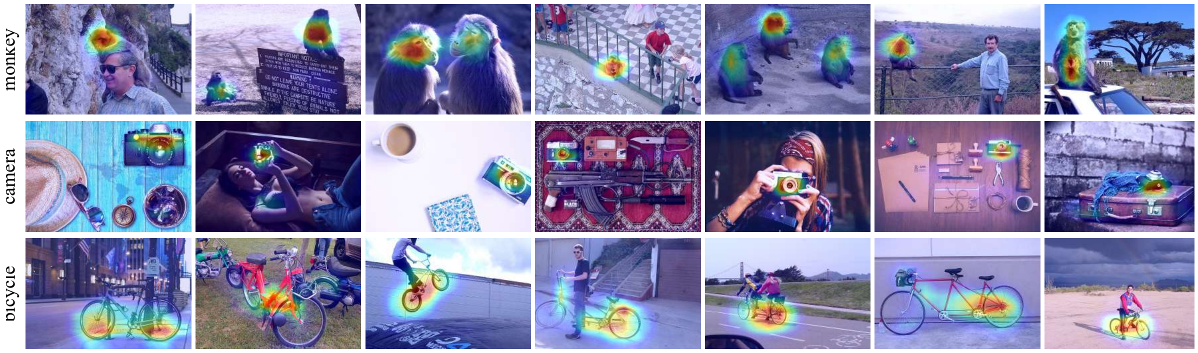

Figure 5. **Visualization of affinity attention maps** learned by GAM using intra-group collaborative learning across all images in each group. Masks are sensitive to co-salient regions with shared attributes, which benefits the consensus representation learning.

derstanding of our GAM module, we visualize the learned attention masks in Figure 5. We find that our global co-attention effectively alleviates the influence of co-occurring noise and focuses on co-salient regions in the image groups, *e.g.*, in both the *monkey* and *bicycle* groups, there are some co-occurring *persons* in some images, but our GAM is not adversely influenced. The global view of GAM enables the most common objects to be detected, while the local pair-wise co-attention cannot distinguish them in the local view.

Effectiveness of GCM. The group collaborating module is designed to enable the consensus *inter-group separability* to distinguish distracting objects from non-common objects. After equip the model with GCM, significant performance improvement (ID-1 versus ID-3) is obtained in Table 1 especially on the challenging CoCA [27] dataset whose images usually contain multiple uncommon and common objects. To investigate the consensus characteristics when the model is trained with the GCM, we visualize the consensus using t-SNE [18] on the CoCA dataset, and compare with the vanilla consensus without the GCM. As shown in Figure 1, the vanilla consensuses (top: CoEGNet [8]) tend to cluster together, even if they belong to different groups, resulting in ambiguous co-saliency detection, especially for objects belonging to similar but different groups. In contrast, the consensuses trained with the GCM (bottom: our method) is more diverse with a higher group variance, for more effective inter-group separability. For quantitative comparison, we evaluate the cosine similarity (↓ lower is better) of the consensus of "guitar" and "violin", and ours (0.32) is much better than CoEGNet (0.75).

Effectiveness of ACM. As shown in Table 1, the classification module introduces better backbone features for the consensus with the auxiliary classification supervision. The ACM improves the baseline performance on all metrics and datasets. This cost-free improvement does not change the network architecture and does not introduce extra computational overhead at inference time, thus has substantial potential to other models and tasks to take advantage of the multi-task learning and more representative features.

4.4. Competing Methods

Since not all CoSOD models have publicly released codes, we only compare our GCoNet with one representative traditional algorithm (CBCD) and five deep-based CoSOD models, including GWD [82], RCAN [43], CSMG [48], GICD [27], and CoEGNet [8]. Following the current state-of-the-art model [27], we also compare with four cutting-edge deep salient object detection (SOD)[7] models: BASNet [28], PoolNet [29], EGNet [30] and SCRN [81]. More complete leaderboard can be found in recent standard benchmark works [8, 74].

[7] SOD methods can also be directly applied to the CoSOD task.

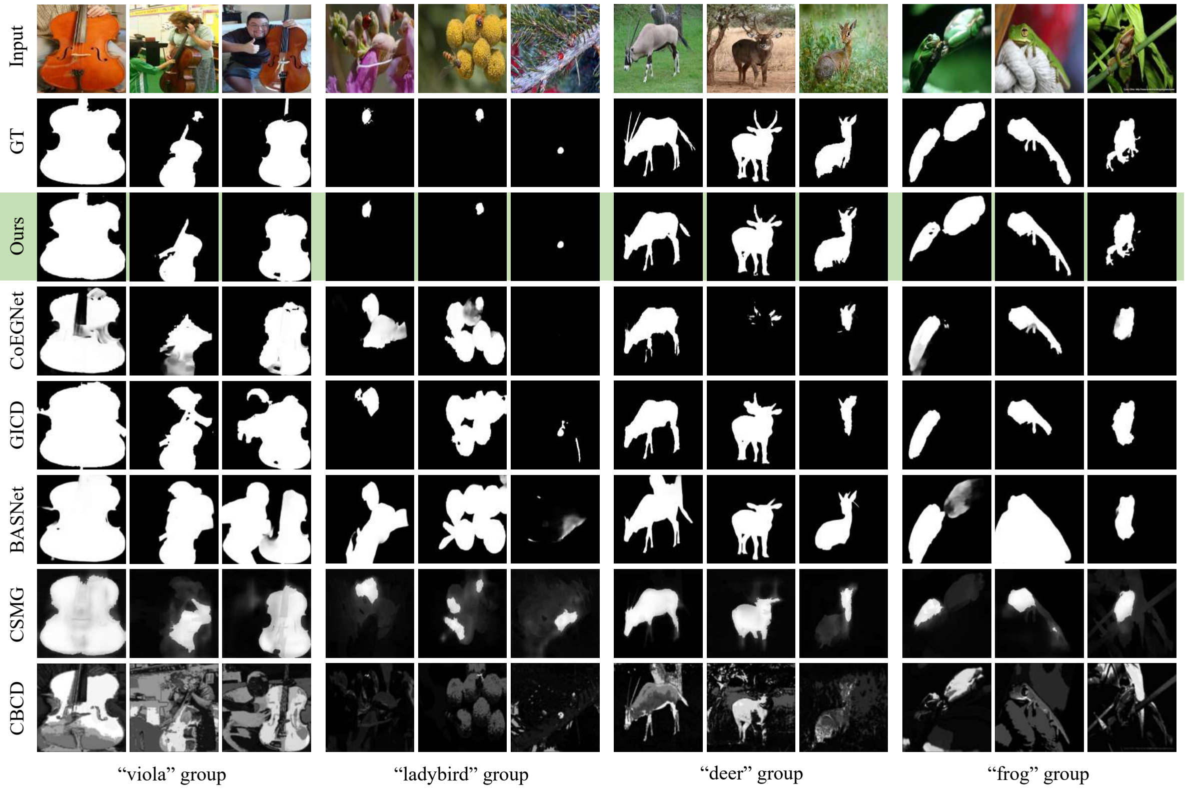

Figure 6. **Qualitative comparisons** of our GCoNet and other methods. "GT" denotes GroundTruth.

Quantitative Results. Table 2 tabulates the quantitative results of our model and state-of-the-art methods. Our model outperforms all of them in all metrics, especially on the challenging CoCA and CoSOD3k datasets. Among these three datasets, CoCA is the most challenging, since the images typically contain other multiple objects in addition to the co-salient objects which are even smaller in size. Our model capitalizes on our better consensus and significantly outperforms other methods especially the SOD methods which are trapped in distinguishing many distracting objects instead. CoSOD3k has similar attributes, and our model still performs much better than other models on this dataset. Cosal2015 is the easiest dataset because its images typically only contain one co-salient object, and therefore the SOD algorithms can easily handle this dataset. Our model cannot take full advantage of the better consensus on this dataset and the improvement is not as significant as on other datasets.

Qualitative Results. Figure 6 shows the saliency maps generated by different methods for qualitative comparison. In these difficult examples, each image contains other multiple objects in addition to the co-salient objects. As aforementioned, the SOD methods can only detect salient objects and fail to distinguish co-salient objects due to their intrinsic limitation. The CoSOD methods perform better than the SOD methods owing to their consensus for distinguishing co-salient regions. However, limited by the their weak consensus, they are still unable to handle the challenging cases. Our model introduces an effective consensus through optimizing intra-group compactness and inter-group separability, and therefore performs much better on detecting co-salient objects.

5. Discussion of Module Cooperation

Our three modules are closely interdependent and mutually reinforced for improving co-saliency detection performance. Combining the GAM and GCM can significantly improve the performance compared to the individual modules. Without the GAM the vanilla consensus is not robust against noise caused by uncommon objects and background, and the low-quality consensus cannot take full advantage of the GCM which heavily relies on the consensus for distinguishing different objects. On the other hand, although the consensus can capture common attributes with the help of the GAM, it is difficult to distinguish different groups without the GCM especially for similar groups. Overall, the GAM produces better consensus with high intra-group compactness to detect co-saliency objects, while the GCM further endows the consensus with inter-group separability for better discriminative ability. Adding ACM, the consensus can benefit from more representative features leveraged by the multi-task learning.

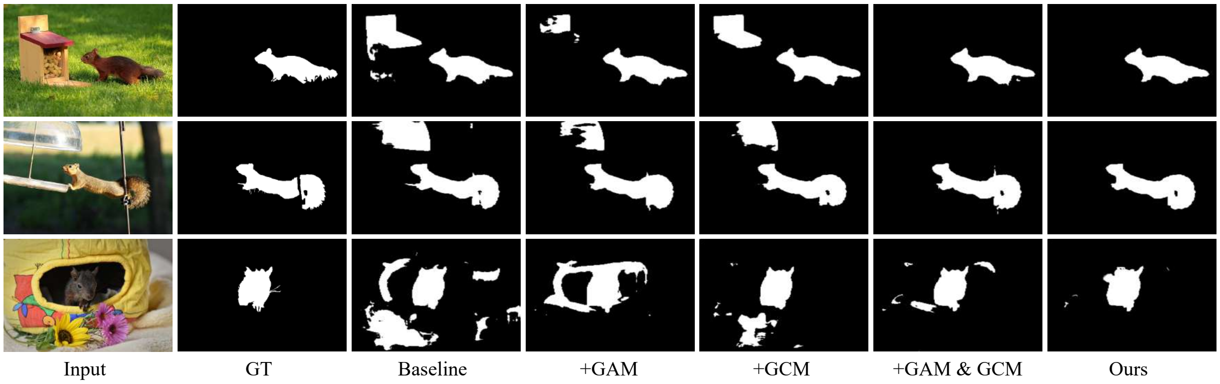

Figure 7. **Qualitative ablation studies** of our GCNet on different modules and their combinations.

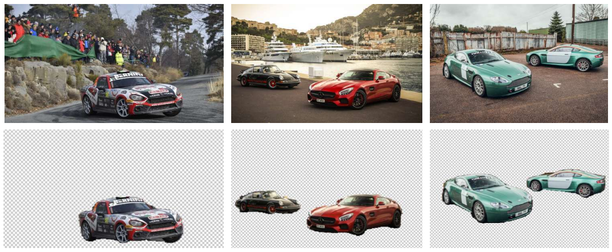

Figure 8. **Application 1.** Content aware object co-segmentation visual results ("GT car") obtained by our GCoNet.

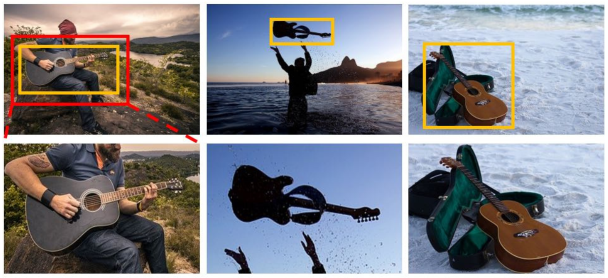

Figure 9. **Application 2.** Co-location based automatic thumbnails ("Raccoon") generated by our GCoNet.

Figure 7 qualitatively analyse their cooperation. The baseline model detects uncommon objects, while the GAM and GCM can slightly ameliorate their adverse influence. When combining the GAM and GCM, the model can effectively capture co-salient objects with the ACM further boosting the co-salient object detection result.

6. Downstream Applications

Here, we show how the extracted co-saliency map can be utilized to generate high-quality segmentation masks for selected closely related downstream image processing tasks.

Application #1: Content-Aware Co-Segmentation. Co-saliency maps have been previously used in pre-processing for unsupervised object segmentation. In our implementation, we first manually select a group of images from the internet by keyword search . Then, co-saliency maps are generated by our GCoNet to automatically mine the salient content of the specific group. Similar to Cheng *et al*. [26], we also utilize GrabCut [83] to obtain the final segmentation results. To initialize GrabCut, we simply choose adaptive threshold [84] to binarize the saliency maps. Figure 8 shows the results of the content-aware object co-segmentation which should benefit existing e-commerce applications requiring background replacement.

Application #2: Automatic Thumbnails. The idea of paired-image thumbnails is derived from the seminal work [11]. With the same goal[8], we present a CNN-based photographic triage application which is valuable for sharing images with friends on the website. As shown in Figure 9, we first generate the yellow box based on the co-saliency map obtained by our GCoNet. Then, we simply enlarge the yellow box to get a larger red box. Finally, we adopt the collection-aware crops technique [11] to produce the results (2^{nd} row).

7. Conclusion

In this paper, we investigate a novel group collaborative learning framework (GCoNet) for CoSOD. We find that group-level consensus can introduce effective semantic information to benefit the representation of both the *intra-group compactness* and *inter-group separability* for CoSOD. Our experiments quantitatively and qualitatively demonstrate the advantage of our GCoNet which outperforms existing state-of-the-art models. In addition, our GCoNet achieves real-time speed (16ms) which can greatly benefit many applications such as co-segmentation, co-localization, and among others.

Acknowledgments

This research was supported in part by Kuaishou Technology, the Research Grant Council of the Hong Kong SAR under grant No. 16201420, and National Natural Science Foundation of China (No. U1936210). We would like to thank the anonymous reviewers for their useful feedbacks.

[8] Note that Jacobs *et al*.'s work [11] is limited to the case of image pairs.

References

[1] 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, 2018.

[2] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *arXiv preprint arXiv:2101.07663*, 2021.

[3] Xuebin Qin, Deng-Ping Fan, Chenyang Huang, Cyril Diagne, Zichen Zhang, Adrià Cabeza Sant'Anna, Albert Suàrez, Martin Jagersand, and Ling Shao. Boundary-aware segmentation network for mobile and web applications. *arXiv preprint arXiv:2101.04704*, 2021.

[4] Tao Zhou, Deng-Ping Fan, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Rgb-d salient object detection: A survey. *CVM*, pages 1–33, 2021.

[5] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. *IEEE TNNLS*, 2021.

[6] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Saleh, Sadegh Aliakbarian, and Nick Barnes. Uncertainty inspired rgb-d saliency detection. *arXiv preprint arXiv:2009.03075*, 2020.

[7] Zuyao Chen, Runmin Cong, Qianqian Xu, and Qingming Huang. Dpanet: Depth potentiality-aware gated attention network for rgb-d salient object detection. *TIP*, 2020.

[8] 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. *TPAMI*, 2021.

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

[10] Xiaochuan Wang, Xiaohui Liang, Bailin Yang, and Frederick WB Li. No-reference synthetic image quality assessment with convolutional neural network and local image saliency. *CVM*, 5(2):193–208, 2019.

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

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

[13] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Deepco3: Deep instance co-segmentation by co-peak search and co-saliency detection. In *CVPR*, 2019.

[14] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *ICCV*, 2019.

[15] Zhifan Gao, Chenchu Xu, Heye Zhang, Shuo Li, and Victor Hugo C de Albuquerque. Trustful internet of surveillance things based on deeply represented visual co-saliency detection. *IoT-J*, 7(5):4092–4100, 2020.

[16] Koteswar Rao Jerripothula, Jianfei Cai, and Junsong Yuan. Efficient video object co-localization with co-saliency activated tracklets. *TCSVT*, 29(3):744–755, 2018.

[17] Koteswar Rao Jerripothula, Jianfei Cai, and Junsong Yuan. Cats: Co-saliency activated tracklet selection for video colocalization. In *ECCV*, 2016.

[18] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008.

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

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

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

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

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

[24] 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*, 2018.

[25] Wenbin Zou, Kidiyo Kpalma, Zhi Liu, and Joseph Ronsin. Segmentation driven low-rank matrix recovery for saliency detection. In *BMVC*, 2013.

[26] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *TPAMI*, 37(3):569–582, 2014.

[27] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In *ECCV*, 2020.

[28] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019.

[29] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, 2019.

[30] 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 *ICCV*, 2019.

[31] Shang-Hua Gao, Yong-Qiang Tan, Ming-Ming Cheng, Chengze Lu, Yunpeng Chen, and Shuicheng Yan. Highly efficient salient object detection with 100k parameters. In *ECCV*, 2020.

[32] Kai Zhao, Shanghua Gao, Wenguan Wang, and Ming ming Cheng. Optimizing the F-measure for threshold-free salient object detection. In *ICCV*, 2019.

[33] Hwann-Tzong Chen. Preattentive co-saliency detection. In *ICIP*, 2010.

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

[35] Kai-Yueh Chang, Tyng-Luh Liu, and Shang-Hong Lai. From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. In *CVPR*, 2011.

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

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

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

[39] 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*, 2019.

[40] Bo Jiang, Xingyue Jiang, Jin Tang, Bin Luo, and Shilei Huang. Multiple graph convolutional networks for co-saliency detection. In *ICME*, 2019.

[41] 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 *CVPR*, 2020.

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

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

[44] Wen-Da Jin, Jun Xu, Ming-Ming Cheng, Yi Zhang, and Wei Guo. Icnet: Intra-saliency correlation network for co-saliency detection. In *NeurIPS*, 2020.

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

[46] 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. *TIP*, 26(7):3196–3209, 2017.

[47] 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. *TIP*, 28(1):56–71, 2018.

[48] 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*, 2019.

[49] 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*, 2018.

[50] 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.

[51] Qijian Zhang, Runmin Cong, Junhui Hou, Chongyi Li, and Yao Zhao. Coadnet: Collaborative aggregation-and-distribution networks for co-salient object detection. In *NeurIPS*, 2020.

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

[53] 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. *TIP*, 28(10):5052–5063, 2019.

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

[55] 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. *TIP*, 27(12):5866–5879, 2018.

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

[57] Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *ICCV*, 2017.

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

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

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

[61] 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.

[62] 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.

[63] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *ECCV*, 2018.

[64] Zihang Lai and Weidi Xie. Self-supervised learning for video correspondence flow. In *BMVC*, 2019.

[65] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, 2019.

[66] Zihang Lai, Erika Lu, and Weidi Xie. Mast: A memory-augmented self-supervised tracker. In *CVPR*, 2020.

[67] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, 2019.

[68] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *CVPR*, 2020.

[69] Zhuwen Li, Qifeng Chen, and Vladlen Koltun. Interactive image segmentation with latent diversity. In *CVPR*, 2018.

[70] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

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

[72] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[73] 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 *CVPR*, 2017.

[74] Deng-Ping Fan, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Huazhu Fu, and Ming-Ming Cheng. Taking a deeper look at the co-salient object detection. In *CVPR*, 2020.

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

[76] John Winn, Antonio Criminisi, and Thomas Minka. Object categorization by learned universal visual dictionary. In *ICCV*, 2005.

[77] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *SSI*, 2021.

[78] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, 2017.

[79] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009.

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

[81] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *ICCV*, 2019.

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

[83] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Interactive foreground extraction using iterated graph cuts. *ACM TOG*, 23:3, 2012.

[84] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: a benchmark and algorithms. In *ECCV*, 2014.