Memory-Aided Contrastive Consensus Learning for Co-salient Object Detection

Peng Zheng[1], Jie Qin[1*], Shuo Wang[2], Tian-Zhu Xiang[3], Huan Xiong[4,5*]
[1] Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2] ETH Zurich, Zurich, Switzerland
[3] Inception Institute of Artificial Intelligence, Abu Dhabi, UAE
[4] Harbin Institute of Technology, China
[5] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{zhengpeng0108, qinjiebuaa, shawnwang.tech, tianzhu.xiang19, huan.xiong.math}@gmail.com

Abstract

Co-salient object detection (CoSOD) aims at detecting common salient objects within a group of relevant source images. Most of the latest works employ the attention mechanism for finding common objects. To achieve accurate CoSOD results with high-quality maps and high efficiency, we propose a novel Memory-aided Contrastive Consensus Learning (MCCL) framework, which is capable of effectively detecting co-salient objects in real time (∼150 fps). To learn better group consensus, we propose the Group Consensus Aggregation Module (GCAM) to abstract the common features of each image group; meanwhile, to make the consensus representation more discriminative, we introduce the Memory-based Contrastive Module (MCM), which saves and updates the consensus of images from different groups in a queue of memories. Finally, to improve the quality and integrity of the predicted maps, we develop an Adversarial Integrity Learning (AIL) strategy to make the segmented regions more likely composed of complete objects with less surrounding noise. Extensive experiments on all the latest CoSOD benchmarks demonstrate that our lite MCCL outperforms 13 cutting-edge models, achieving the new state of the art (∼5.9% and ∼6.2% improvement in S-measure on CoSOD3k and CoSal2015, respectively). Our source codes, saliency maps, and online demos are publicly available at https://github.com/ZhengPeng7/MCCL.

Introduction

Co-salient object detection (CoSOD) aims at detecting the most common salient objects among a group of source images. Compared with the standard salient object detection (SOD) task, CoSOD is more challenging for distinguishing co-occurring objects across a group of images where salient objects of different classes and non-salient objects of the same class are both hard distractors. CoSOD methods also show their advantage of acting as a pre-processing step for other computer vision tasks, such as semantic segmentation (Zeng et al. 2019), co-segmentation (Hsu, Lin, and Chuang 2019), and object tracking (Zhou et al. 2021), *etc*..

Previous works tend to exploit various types of consistency among the image group to solve the CoSOD task, including shared features (Fu, Cao, and Tu 2013) and common semantic information (Han et al. 2018). With the success of unified models in up-stream tasks (Ren et al. 2015; Xiao et al. 2017), the latest CoSOD models try to address salient object detection and common object detection in a unified framework (Fan et al. 2021, 2022; Zhang et al. 2020c). Despite the promising performance achieved by these methods, most of them only focus on learning better consistent feature representations in an individual group (Zhang et al. 2020c; Wei et al. 2017; Zhang et al. 2020b; Cong et al. 2022; Jin et al. 2020; Tang et al. 2022), which may make them suffer from the following limitations. First, images from the

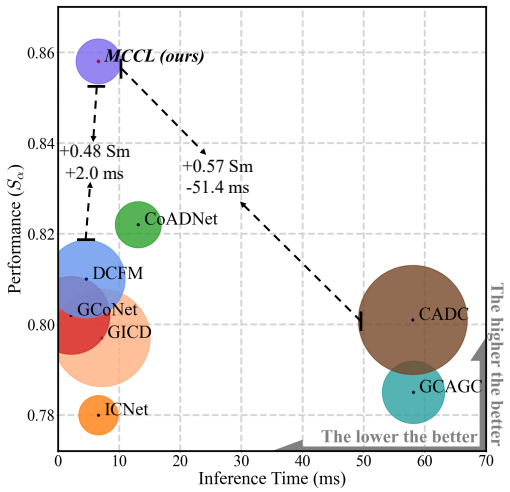

Figure 1: Accuracy (S-measure) and inference time (ms) of the update-to-date and representative deep-learning-based CoSOD methods on the CoSOD3k dataset (Fan et al. 2022). We conduct the comparison on both accuracy (the vertical axis) and speed (the horizontal axis) among seven existing representative CoSOD models and our MCCL. Larger bubbles mean larger volume of the model weights. Our MCCL achieves great performance (0.860 S-measure) in real time (7.6 ms) with a lightweight model (104.5 Mb weights and 5.93G FLOPs). All the methods are tested with batch size 2 on one A100-80G, an online benchmark for inference speed can be found at: https://github.com/ZhengPeng7/CoSOD_fps_collection.

*Corresponding authors.

same group can only act as positive samples of each other. Consensus representations learned from all positive samples might be difficult to be distinguished due to the lack of inter-group separability. Besides, the number of images in a single group is usually insufficient for models to learn robust and unique representations, which can be easily distinguishable from others. Due to the higher complexity of image context in real-world applications, the number of object classes will increase significantly, making the consensuses have a higher risk of getting closer to others and being hard to identify. In this situation, a module designed for bridging the cross-group connection and learning consensuses of distinction is in high demand.

To achieve accurate and fast CoSOD, we propose Memory-aided Contrastive Consensus Learning (MCCL), which exploits common features within each group and identifies distinctions among different groups, guiding the model to produce co-saliency maps with high integrity. To fulfill the above goal, three key components are proposed in MCCL. **First**, we present the Group Consensus Aggregation Module (GCAM) for mining the common feature within the same group by the correlation principle. **Second**, we introduce the Memory-based Contrastive Module (MCM) to conduct robust contrastive learning with a long-term memory. More concretely, the consensus of each class is saved and updated with momentum in a memory queue to avoid the instability of online contrastive learning. **Third**, we employ the Adversarial Integrity Learning (AIL) to improve the integrity and quality of predicted maps in an adversarial fashion, where a discriminator identifies whether the masked regions are obtained from predicted or ground-truth maps. Analogous to generative adversarial networks (Goodfellow et al. 2014), our model tries to fool the discriminator and produce high-quality and high-integrity maps that can mask complete and intact objects.

Our main contributions can be summarized as follows:

- We establish a fast yet strong CoSOD baseline with the Transformer, which outperforms most existing methods that are sophisticatedly equipped with many components.
- We introduce the Group Consensus Aggregation Module (GCAM) to generate the consensus of each group in an effective way. To make the consensus more discriminative to each other, we propose the Memory-based Contrastive Module (MCM) in a metric learning way.
- Furthermore, the Adversarial Integrity Learning (AIL) is proposed to improve the quality and integrity of predicted co-saliency maps in an adversarial learning manner.
- We conduct extensive experiments to validate the superiority of our MCCL. Extensive quantitative and qualitative results show that our MCCL can outperform existing CoSOD models by a large margin.

Related Work

Salient Object Detection

Before the deep learning era, handcrafted features played the most critical role in detection (Cheng et al. 2011; Jiang et al. 2013; Li et al. 2013) among the traditional SOD methods. When it comes to the early years of deep learning, features are usually extracted from image patches, which will then be used to generate object proposals (Wang et al. 2015; Zhang et al. 2016b; Kim and Pavlovic 2016), or super-pixels (Li and Yu 2015; Zhao et al. 2015) as processing units. As stated in (Liu et al. 2022), the network architectures of existing SOD methods can be divided into five categories, *i.e.*, U-shape, side-fusion, multi-branch, single-steam, and multi-stream. So far, U-shape has been the most widely used architecture (Ronneberger, Fischer, and Brox 2015), especially when the fusion between low-level and high-level features is needed. Supervision on the multi-stage output is also employed at the early stages by aggregating features from different levels of networks in the U-shape architecture to make the output features more robust and stable (Zhao et al. 2019; Fan et al. 2021; Zhang et al. 2020c). (Zhang et al. 2018; Liu, Han, and Yang 2018; Zhao and Wu 2019) employed the attention mechanism in their models for further improvement. Besides, some external information is also introduced as extra guidance for training, such as boundary (Qin et al. 2019), edge (Zhao et al. 2019), and depth (Zhao et al. 2019).

Co-salient Object Detection

CoSOD emphasizes on detecting salient objects across groups of images rather than in a single image. Traditional CoSOD methods utilize handcrafted cues (*e.g.*, superpixels (Achanta et al. 2012)) to explore the correspondence of images. In contrast, the deep learning-based methods learn the consensus feature representation of common objects in an end-to-end manner (Wei et al. 2017; Han et al. 2018). Various model architectures are applied to improve the CoSOD performance, including CNN-based (Fan et al. 2021; Zhang et al. 2020b,c) and Transformer-based models (Tang 2021). Though some of the existing methods investigate both intra-group and inter-group cues (Fan et al. 2021), there is still much room for improvement in the fully coordinated and simultaneous use of intra-group and inter-group information.

Integrity Learning for Saliency Maps

The quality of saliency maps has attracted much attention in recent years to make existing saliency-related tasks closer to real-world applications. (Li and Yu 2016) tries to guide their models to learn integrity via the collaboration between global context and local objects. TSPOANet (Liu et al. 2019) adopts a capsule network to model the part-object relationship to achieve better wholeness and uniformity of segmented salient objects. In (Qin et al. 2019), a hybrid loss is applied for more focus on improving the boundary of predicted maps. Furthermore, (Zhuge et al. 2022) investigates more into the integrity issue in SOD and tries solving this problem with their carefully designed components. In (Zheng et al. 2022), a confidence enhancement module is proposed to make the predicted maps more binarized.

Methodology

In this section, we first introduce the overall architecture of our MCCL for the CoSOD task. Then, we sequentially introduce the proposed three key components: Group Consensus

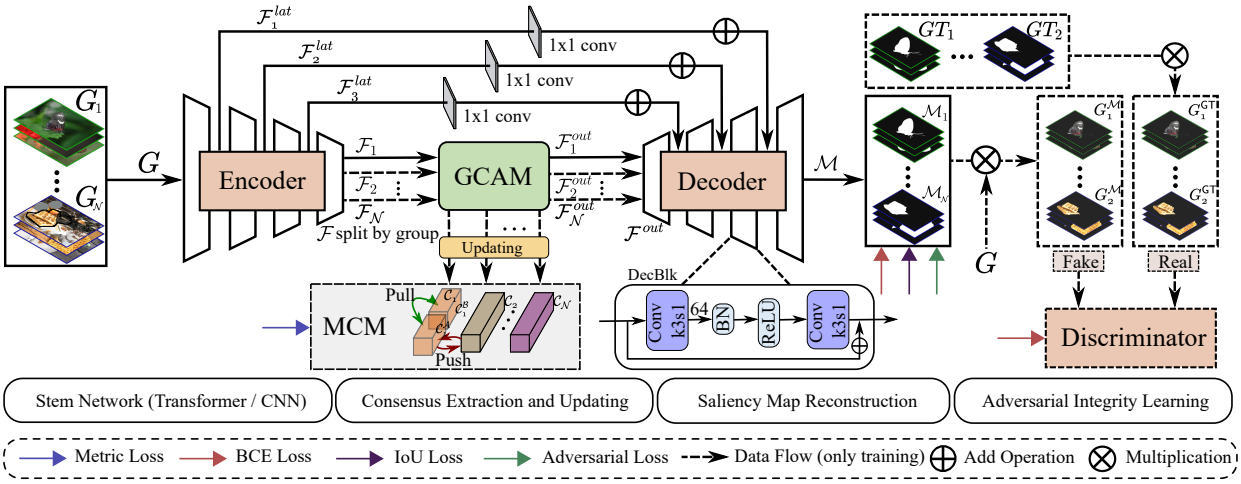

Figure 2: Overall framework of the proposed Memory-aided Contrastive Consensus Learning (MCCL). Input images are obtained from multiple groups and fed to an encoder. First, we employ the Group Consensus Aggregation Module (GCAM), where the intra-group features of each group can be learned separately. With the consensus learned from each single group, the consensus features are updated in the memory of each class in the queue of Memory-based Contrastive Module (MCM). Then, contrastive learning is conducted to make the consensus more discriminative to each other. Each stage of our encoder and decoder is connected with only a 1x1 convolution layer for feature adding with the least computation. Our decoder is composed of four DecBlk, which is the vanilla residual block. We design our model as simple as possible to make our study more open and solid. Finally, saliency maps of all groups are predicted based on the supervision from the BCE and IoU losses.

Aggregation Module (GCAM), Memory-based Contrastive Module (MCM), and Adversarial Integrity Learning (AIL).

First, GCAM is used to exploit the common features of images in the same group. Second, MCM is applied to make the learned consensus of different groups more robust and discriminative to each other. Finally, we employ AIL to improve the integrity and quality of predicted maps in an adversarial way. Note that MCM and AIL are only used during training and thus can be entirely discarded during inference for a more lightweight model.

Overview

Fig. 2 illustrates the basic framework of the proposed MCCL including the learning pipeline. Different from existing CoSOD models that take images from a single group (Fan et al. 2022; Zhang et al. 2020b; Jin et al. 2020; Zhang et al. 2020c) as input, our model receives images from multiple groups as input, bringing the potential to bridge the intersection between different groups.

First, we take images of N (default as 2) groups as the input $\{G_1, G_2, ..., G_N\}$. We concatenate all the images as a whole batch G, which is then fed to the encoder. With the backbone network (default as the Transformer network PVTv2 (Wang et al. 2022)) as our encoder, embedded features are extracted as $\mathcal{F}$, which is then split by their group categories as $\{\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_N\}$, where $\mathcal{F}_N = \{F_{N,s}\}_{s=1}^{S} \in \mathbb{R}^{S\times C\times H\times W}$, C denotes the channel number, $H \times W$ means the spatial size, and N is the group size. Meanwhile, the intermediate features $\{\mathcal{F}_1^{lat}, \mathcal{F}_2^{lat}, \mathcal{F}_3^{lat}\}$ at different stages of our encoder are saved and fed to their corresponding stages of the decoder by a 1x1 convolutional layer.

Then, $\{\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_N\}$ are sequentially fed to GCAM to obtain the consensus of each group. With the consensus of groups $\{\mathcal{F}_1^{out}, \mathcal{F}_2^{out}, ..., \mathcal{F}_N^{out}\}$, the memory of the corresponding classes is updated in the queue with momentum, supervised by a metric learning loss used in (Zheng et al. 2022).

Furthermore, the consensus of all groups is concatenated as $\mathcal{F}^{out}$ and fed to the decoder, which consists of four stacked standard residual blocks and combines the early features from lateral connections. Co-saliency maps $\mathcal{M}$ are generated at the end of the decoder. The prediction of the decoder M is supervised by the Binary Cross Entropy (BCE) loss and the Intersection over Union (IoU) loss, which provide pixel-level and region-level supervision, respectively.

Finally, the predicted co-saliency maps $\mathcal{M}$ are facilitated together with the source images G and the ground-truth maps GT. The pixel-wise multiplication between G and M leads to $G^{\mathcal{M}}$, and we obtain G^{GT} in a similar way. $G^{\mathcal{M}}$ and G^{GT} are then fed to an independent discriminator, identifying whether the masked images are generated by the ground-truth maps G^{GT} or the source images G, which include intact and complete objects. Accordingly, the adversarial loss from the discriminator is applied to the whole generator, and the BCE loss is given to the discriminator.

Group Consensus Aggregation Module

In the wild, objects of the same category tend to share similar appearance, which has been exploited in many related tasks, such as video tracking (Wang, Jabri, and Efros 2019) and semantic segmentation (Zhang et al. 2019), where the correspondence between common objects is used as

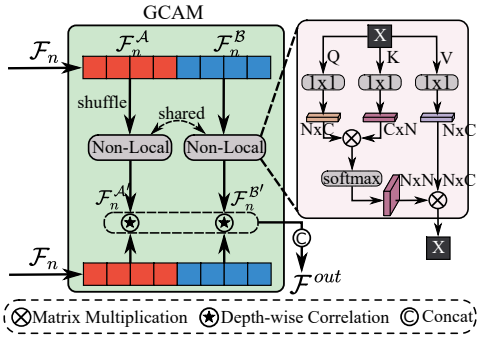

Figure 3: Group Consensus Aggregation Module. The feature of the encoder is fed to the GCAM and handled group by group. The original features of one group are evenly split and shuffled before being fed to the non-local block. The depth-wise correlation bridges the semantic interaction between the consensus and original features.

prior information. Here we also apply this mechanism to CoSOD. Similar to (Fan et al. 2021), we employ the non-local block (Wang et al. 2018) to extract the affinity feature. As shown in Fig. 3, we first split the output feature of the encoder $\mathcal{F}_n$ into $\{\mathcal{F}_n^A, \mathcal{F}_n^B\}$, which are then shuffled and fed into the non-local block. Subsequently, in the non-local block, we compute the affinity map of the feature and conduct matrix multiplication between the affinity map and the value feature (*i.e.*, 'V' in the non-local block) to obtain the consensus feature $\{\mathcal{F}_n^{A'}, \mathcal{F}_n^{B'}\}$. Finally, we perform depth-wise correlation to fuse the original feature with the consensus feature, and concatenate them to form the final consensus representation $\mathcal{F}_{out}$.

Memory-Based Contrastive Module

Metric learning is a widely-used technique that contributes to distinguishing features of different clusters and works in many tasks, including CoSOD (Han et al. 2018; Zhang, Meng, and Han 2017; Zheng et al. 2022). However, CoSOD datasets only contain a limited number of images (tens of images) of limited groups (less than 300 groups). In such cases, naive metric learning cannot work very well due to the small number of samples insufficient for distance measurement.

To overcome this issue, some contrastive learning methods introduce the memory queue to achieve more robust contrastive learning with a long-term memory, such as MoCo (He et al. 2020), OIM (Xiao et al. 2017), *etc*.. Inspired by these works, we propose the MCM, which saves the consensus features of each class into memory blocks and updates the corresponding blocks with momentum in every batch. To be more specific, as shown in Fig. 2, the consensuses of all groups $\{\mathcal{F}_1^{out}, \mathcal{F}_2^{out}, ..., \mathcal{F}_N^{out}\}$ are saved or updated in their own memory blocks as $\{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_N\}$. The memory update is as follows:

$$\mathcal{C}_1^t = \beta * \mathcal{C}_1^{t-1} + (1-\beta) * \mathcal{F}_1^{out}, \tag{1}$$

where β denotes the momentum factor and is set to 0.1 by default. When β is set to 0, the MCM belongs to fully online metric learning.

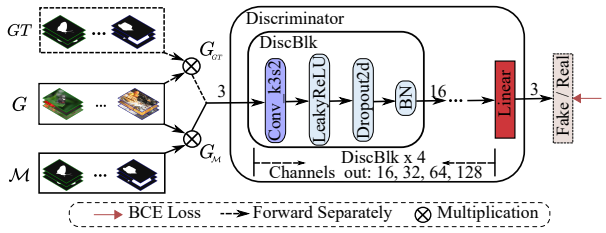

Figure 4: Discriminator used in our AIL. The discriminator has four discrimination blocks (DiscBlk) stacked sequentially, with 16, 32, 64, and 128 as the number of their output channels, respectively. Note that there is no batch normalization layer in the first DiscBlk in our implementation.

As demonstrated in the MCM, each memory block splits itself into two parts, $\mathcal{C}_1^A$ and $\mathcal{C}_1^B$. In this case, $\mathcal{C}_1^B$ is viewed as the positive samples of $\mathcal{C}_1^A$, and the whole $\mathcal{C}_2$ is considered as the negative samples of $\mathcal{C}_1^A$ (Zheng et al. 2022). Then, the loss of MCM can be computed by the GST loss (Zheng et al. 2022) as follows:

$$L_{\text{Triplet}}(C_1, C_2) = ||\mathcal{C}_1^A - \mathcal{C}_1^B||_2 - ||\mathcal{C}_1^A - \mathcal{C}_2^B||_2 + \alpha \tag{2}$$

$$L_{\text{MCM}} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} L_{\text{Triplet}}(C_i, C_j), \tag{3}$$

where α denotes the margin used in the triplet loss (Schroff, Kalenichenko, and Philbin 2015), which is set to 0.1. $||\cdot||_2$ means the l_2 norm of the input.

Adversarial Integrity Learning

Although some latest works have investigated the integrity of SOD (Zhuge et al. 2022), they try to solve this problem by designing sophisticated model architectures and critical components to make predicted maps of higher integrity. These attempts may lead to maps with better quality, but the motivation of their design is not intuitive to the integrity problem.

To explicitly solve this problem, we propose the Adversarial Integrity Learning (AIL) in our framework. Three data sources exist in AIL, *i.e.*, the source images, the ground-truth maps, and the predicted maps in the current batch. During training, we perform pixel-wise multiplications on two pairs, *i.e.*, (source images, ground-truth maps) and (source images, predicted maps), as shown in Fig. 4, to obtain G_{GT} and G_M, respectively. Then, we employ a discriminator to identify whether the regions of source images masked by these two maps are real or fake, as shown in Fig. 2. Obviously, the regions masked by the ground-truth maps are complete and intact objects with 100% integrity. During training, the loss from the discriminator guides the generator to produce maps that can localize objects with higher accuracy and integrity. The ablation results are shown in Fig. 7.

Objective Function

As shown in Fig. 2, the objective function L_{sal} of the main network (generator) is a weighted combination of the low-level losses (*i.e.*, BCE and IoU losses) and the high-level

losses (metric and adversarial losses). And the discriminator involves the BCE loss. The details of L_{MCM} can be found in the 'Methodology' section above. The BCE, IoU, and adversarial losses are as follows:

$$L_{\mathrm{BCE}} = -\sum [Y log(\hat{Y}), (1-Y) log(1-\hat{Y})], \quad (4)$$

$$L_{\mathrm{IoU}} = 1 - \frac{1}{N}\sum \frac{Y \cap \hat{Y}}{Y \cup \hat{Y}}, \quad (5)$$

where Y is the ground-truth map, $\hat{Y}$ is the predicted map.

$$\hat{T} = discriminator(\hat{Y} \cdot G), \quad (6)$$

$$L_{\mathrm{adv}} = -\sum [T log(\hat{T}), (1-T) log(1-\hat{T})], \quad (7)$$

where $\hat{Y}$ is the predicted map, G denotes the source images, $\cdot$ denotes the pixel-wise multiplication, $\hat{T}$ and T denote the prediction of discriminator on whether $\hat{Y}$ and Y is the ground-truth map, respectively.

Therefore, our final objective function is:

$$L_{\mathrm{sal}} = \lambda_1 L_{\mathrm{BCE}} + \lambda_2 L_{\mathrm{IoU}} + \lambda_3 L_{\mathrm{MCM}} + \lambda_4 L_{\mathrm{adv}}, \quad (8)$$

$$L_{\mathrm{disc}} = \lambda_5 L_{\mathrm{BCE}}, \quad (9)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, and λ_5 are respectively set to 30, 0.5, 3, 10, and 3 to keep all the losses at a reasonable scale at the beginning of training to benefit the optimization.

Experiments

Datasets

Training Sets. We follow (Zhang et al. 2020b) to use DUTS_class (Zhang et al. 2020c) and COCO-SEG (Wang et al. 2019) as our training sets. The whole DUTS_class is divided into 291 groups, which contain 8,250 images in total. COCO-SEG contains 200k images of 78 groups and corresponding binary maps.

Test Sets. For a comprehensive evaluation of our MCCL, we test it on three widely used CoSOD datasets, *i.e.*, CoCA (Zhang et al. 2020c), CoSOD3k (Fan et al. 2022), and CoSal2015 (Zhang et al. 2016a). Among the three datasets, CoCA is the most challenging one. It is of much higher diversity and complexity in terms of background, occlusion, illumination, surrounding objects, *etc*.. Following the latest benchmark (Fan et al. 2022), we do not evaluate on iCoseg (Batra et al. 2010) and MSRC (Winn, Criminisi, and Minka 2005), since only one salient object is given in most images there. It is more convincing to evaluate CoSOD methods on images with more salient objects, which is closer to real-life applications.

Evaluation Protocol

Following GCoNet (Fan et al. 2021), we employ the S-measure (Fan et al. 2017), maximum F-measure (Achanta et al. 2009), maximum E-measure (Fan et al. 2018), and mean absolute error (MAE) to evaluate the performance in our experiments.

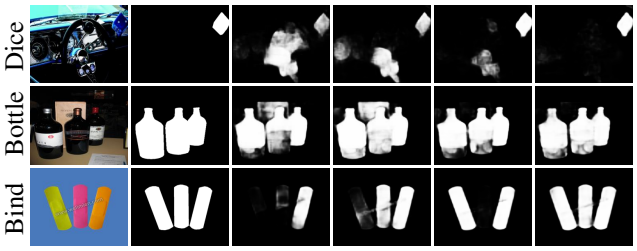

Figure 5: Qualitative ablation studies of different modules and their combinations in our MCCL. (a) Source image; (b) Ground truth; (c) w/ GCAM; (d) w/ GCAM+MCM; (e) w/ GCAM+AIL; (f) w/ GCAM+MCM+AIL, *i.e.*, the full version of our model.

Implementation Details

We select samples from DUTS_class (Zhang et al. 2020c) and COCO-SEG (Wang et al. 2019) alternatively, and set the batch size as follows:

$$batchsize = min(\#group1, ..., \#groupN, 48), \quad (10)$$

where $\#$ means the image number in the corresponding group. The images are resized to 256x256 for training and inference. The output maps are resized to the original size for evaluation. We apply three data augmentation strategies, *i.e.*, horizontal flip, color enhancement, and rotation. Our MCCL is trained for 250 epochs with the AdamW optimizer (Loshchilov and Hutter 2019). The initial learning rate is 1e-4 and divided by 10 at the last 20th epoch. The whole training process takes $\sim$3.5 hours and consumes only $\sim$7.5GB GPU memory. All the experiments are implemented based on the PyTorch library (Paszke et al. 2019) with a single NVIDIA RTX3090 GPU.

Ablation Study

We conduct the ablation study to validate the effectiveness of each component (*i.e.*, GCAM, MCM, and AIL) employed in our MCCL. The qualitative results regarding each module and the combination are shown in Fig. 5.

Baseline. We establish a solid CoSOD network as the baseline. To keep pace with the latest Transformer network, we also build our model with both Transformer and convolutional neural networks as the backbone. Following GCoNet (Fan et al. 2021), we feed images of multiple classes and their ground-truth as the input to train our MCCL. Compared with previous CoSOD models, our baseline network achieves promising performance with a simpler architecture and much higher speed. To be consistent with the widely used Transformer network (Dosovitskiy et al. 2021; Wang et al. 2021; Wang et al. 2022), in contrast with the previous CoSOD models (Zhang et al. 2020c; Fan et al. 2021; Wu, Su, and Huang 2019; Yu et al. 2022), we make our model shallower and build it with only four stages in its encoder and decoder. To achieve a pure and fast baseline network, we firstly substitute all the complex blocks in each lateral connection used in (Zhang et al. 2020c; Fan et al. 2021; Wu, Su, and Huang 2019; Yu et al. 2022) with a single

Method	Pub.	CoCA (Zhang et al. 2020c)				CoSOD3k (Fan et al. 2022)				CoSal2015 (Zhang et al. 2016a)			
		$E_\xi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$	$E_\xi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$	$E_\xi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$
CBCS (2013)	TIP	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 (2017)	IJCAI	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 (2019)	IJCAI	0.702	0.616	0.422	0.160	0.808	0.744	0.688	0.130	0.842	0.779	0.764	0.126
GCAGC (2020a)	CVPR	0.754	0.669	0.523	0.111	0.816	0.785	0.740	0.100	0.866	0.817	0.813	0.085
GICD (2020c)	ECCV	0.715	0.658	0.513	0.126	0.848	0.797	0.770	0.079	0.887	0.844	0.844	0.071
ICNet (2020)	NeurIPS	0.698	0.651	0.506	0.148	0.832	0.780	0.743	0.097	0.900	0.856	0.855	0.058
CoADNet (2020b)	NeurIPS	-	-	-	-	0.878	0.824	0.791	0.076	0.914	0.861	0.858	0.064
DeepACG (2021)	CVPR	0.771	0.688	0.552	0.102	0.838	0.792	0.756	0.089	0.892	0.854	0.842	0.064
GCoNet (2021)	CVPR	0.760	0.673	0.544	0.105	0.860	0.802	0.777	0.071	0.887	0.845	0.847	0.068
CoEGNet (2022)	TPAMI	0.717	0.612	0.493	0.106	0.837	0.778	0.758	0.084	0.884	0.838	0.836	0.078
CADC (2019)	ICCV	0.744	0.681	0.548	0.132	0.840	0.801	0.859	0.096	0.906	0.866	0.862	0.064
DCFM* (2022)	CVPR	0.783	0.710	**0.598**	**0.085**	0.874	0.810	0.805	0.067	0.892	0.838	0.856	0.067
UFO (2022)	arXiv	0.782	0.697	0.571	0.095	0.874	0.819	0.797	0.073	0.906	0.860	0.865	0.064
Ours	Sub.	**0.796**	**0.714**	0.590	0.103	**0.903**	**0.858**	**0.837**	**0.061**	**0.927**	**0.890**	**0.891**	**0.051**

Table 1: Quantitative comparisons between our MCCL and other methods. "↑" ("↓") means that the higher (lower) is better. * denotes the state-of-the-art method. UFO (Su et al. 2022) is still an arXiv paper and does not show much improvement compared with DCFM (Yu et al. 2022), so we set DCFM as the previous SoTA.

Modules			CoCA (Zhang et al. 2020c)				CoSOD3k (Fan et al. 2022)				CoSal2015 (Zhang et al. 2016a)			
GCAM	MCM	AIL	$E_\xi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$	$E_\xi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$	$E_\xi^{\max}\uparrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$\epsilon\downarrow$
			0.756	0.683	0.553	0.118	0.880	0.828	0.798	0.075	0.905	0.866	0.861	0.062
✓			0.779	0.709	0.577	0.103	0.894	0.851	0.824	0.061	0.921	0.884	0.882	0.053
✓	✓		0.788	0.711	0.585	0.100	0.898	0.853	0.828	0.060	0.925	0.889	0.886	**0.050**
✓		✓	0.789	0.714	0.585	**0.097**	0.900	0.855	0.831	0.060	0.924	0.888	0.887	0.053
✓	✓	✓	**0.796**	**0.714**	**0.590**	0.103	**0.903**	**0.858**	**0.837**	**0.061**	**0.927**	**0.890**	**0.891**	0.051

Table 2: Quantitative ablation studies of the proposed components in our MCCL. The components include the GCAM, MCM, AIL, and their combinations.

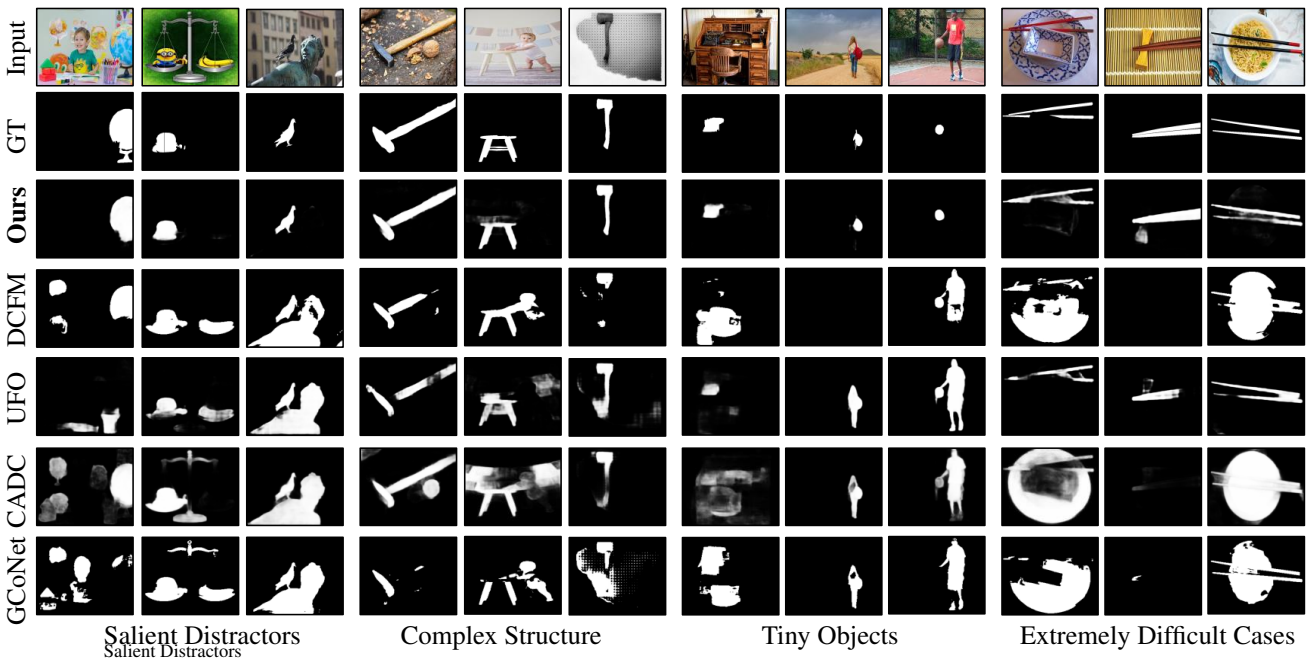

Figure 6: Qualitative comparisons of our MCCL and other methods. 'GT' denotes the ground truth. We select the results of hard cases due to various reasons. The 'Extremely Difficult Cases' means the chopstick group on the test set of CoCA (Zhang et al. 2020c), as chopsticks are thin, tall, and hard to detect. This could be the most difficult case among all groups on the existing test sets.

CoSOD Models	AIL	CoCA (Zhang et al. 2020c)			
GCoNet (2021)		0.760	0.673	0.544	0.105
GCoNet (2021)	✓	0.777	0.681	0.549	0.106
GICD (2020c)		0.715	0.658	0.513	0.126
GICD (2020c)	✓	0.718	0.675	0.524	0.127

Table 3: Quantitative ablation studies of the proposed AIL on different models. We apply the proposed AIL to other CoSOD models (Fan et al. 2021; Zhang et al. 2020c) and conduct the evaluation on CoCA (Zhang et al. 2020c).

1x1 convolution layer (Lin et al. 2017) as the vanilla Feature Pyramid Network (FPN) (Lin et al. 2017) does. Secondly, we set only a single residual block as the decoding block, where the output is added with the features from the lateral connection. Finally, instead of the multi-stage supervision on all stages of the decoder (Zhang et al. 2020c; Fan et al. 2021; Yu et al. 2022), we set the pixel-level supervision on only the final output with a weighted sum of BCE and IoU losses to guide the model locally and globally, respectively. Our baseline can beat most of the existing CoSOD methods, and thus can be referred to as a strong baseline for others in the future.

GCAM. As the performance evaluated in Tab. 2, our GCAM brings much improvement not only on CoCA and CoSOD3k that focus more on complex context with multiple objects, but also on CoSal2015 that is a relatively simple dataset but needs more attention on the precise SOD in a simple environment.

MCM. In Tab. 2, MCM shows its consistent improvement in all metrics on all datasets. As shown in Fig. 5, MCM helps our model make more accurate predictions than those models without it.

AIL. AIL guides our model to learn the integrity of predicted regions, and the produced co-saliency maps tend to be more robust and contain one or more complete and intact objects. As shown in Fig. 7, the improvement brought by AIL can be seen from three perspectives. 1) On the object level, AIL increases the completeness of predicted maps of slender objects, which is usually hard to be fully detected. 2) Inside the object, AIL helps fill the unconfident regions which break the structural integrity of the detected objects. 3) Outside the object, AIL suppresses the distractors, with which the regions include just not a complete object.

In summary, 1) GCAM splits features into two parts, accelerating the affinity generation. 2) MCM is initially motivated by OIM (Xiao et al. 2017). Differently, MCM saves features by classes instead of identities; OIM uses the final normalized features to update the queue, while we use the consensus generated by GCAM, which aligns well with CoSOD. 3) The adversarial learning strategy in AIL can also be found in domain adaptation (Ganin et al. 2016), but we are the first to accommodate it to SOD and to use the segmented regions for discrimination. We also apply AIL to other CoSOD models to validate its high generalizability, as shown in Tab. 3.

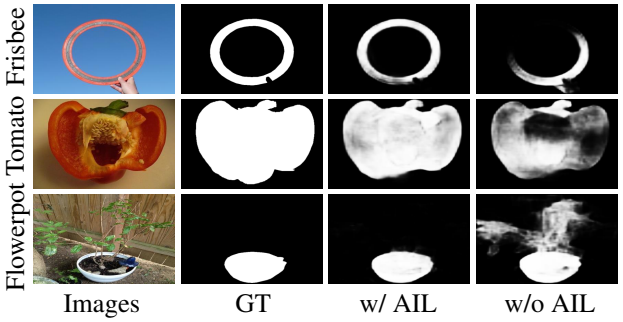

Figure 7: Qualitative ablation studies of our AIL. We conduct the qualitative comparison between the baseline model with (w/) and without (w/o) the proposed AIL.

Comparison to State-of-the-Art Methods

To make a comprehensive comparison, we compare our MCCL with one traditional classical algorithm CBCS (Fu, Cao, and Tu 2013) and 12 update-to-date deep learning based CoSOD models (see Tab. 1 for all methods used for comparison). Since CoSOD methods have gained much improvement in the past few years and obtained much better performance compared with single-SOD methods, we do not list the single-SOD ones, following previous works (Fan et al. 2022, 2021; Zhang et al. 2020c; Yu et al. 2022). The detailed leaderboard of previous methods can be found in (Fan et al. 2022).

Quantitative Results. Tab. 1 shows the quantitative results of our MCCL and previous competitive methods. Given the above results, we can see that our MCCL outperforms all the existing methods, especially on the CoSOD3k (Fan et al. 2022) and CoSal2015 (Zhang et al. 2016a) datasets, where the ability to detect salient objects is in a higher priority than finding objects with the common class.

Qualitative Results. Fig. 6 shows the co-saliency maps predicted by different methods for a clear qualitative comparison, where we provide four different types of complex samples from CoCA (Zhang et al. 2020c) and CoSOD3k (Fan et al. 2022). Compared with existing models, our MCCL shows a stronger ability to eliminate distractors, detect tiny targets, and handle the objects that blend into complex scenes. The extremely difficult cases in which other up-to-date methods fail most of the time further demonstrate the more robust performance of our MCCL.

Conclusion

In this paper, we investigate a novel memory-aided contrastive consensus learning framework (*i.e.*, MCCL) for CoSOD. As experiments show, the memory-based contrastive learning with group consensus works effectively to enhance the representation capability of the obtained group features. Besides, the adversarial integrity learning strategy does benefit the saliency model, with the potential to improve the integrity and quality of saliency maps for a variety of SOD and CoSOD models in a general way.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. 62276129) and the Natural Science Foundation of Jiangsu Province (No. BK20220890).

References

Achanta, R.; Hemami, S.; Estrada, F.; and Susstrunk, S. 2009. Frequency-tuned salient region detection. In *Conf. Comput. Vis. Pattern Recog.*, 1597–1604.

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11): 2274–2282.

Batra, D.; Kowdle, A.; Parikh, D.; Luo, J.; and Chen, T. 2010. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *Conf. Comput. Vis. Pattern Recog.*, 3169–3176.

Cheng, M.-M.; Zhang, G.-X.; Mitra, N. J.; Huang, X.; and Hu, S. 2011. Global contrast based salient region detection. In *Conf. Comput. Vis. Pattern Recog.*, 409–416.

Cong, R.; Yang, N.; Li, C.; Fu, H.; Zhao, Y.; Huang, Q.; and Kwong, S. 2022. Global-and-local collaborative learning for co-Salient object detection. *IEEE Trans. Cybern.*, 1–1.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Int. Conf. Learn. Represent.*

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-Measure: A New Way to Evaluate Foreground Maps. In *Int. Conf. Comput. Vis.*, 4558–4567.

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *Int. Joint Conf. Artif. Intell.*, 698–704.

Fan, D.-P.; Li, T.; Lin, Z.; Ji, G.-P.; Zhang, D.; Cheng, M.-M.; Fu, H.; and Shen, J. 2022. Re-thinking Co-Salient Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1–1.

Fan, Q.; Fan, D.-P.; Fu, H.; Tang, C.-K.; Shao, L.; and Tai, Y.-W. 2021. Group Collaborative Learning for Co-Salient Object Detection. In *Conf. Comput. Vis. Pattern Recog.*, 12283–12293.

Fu, H.; Cao, X.; and Tu, Z. 2013. Cluster-based co-saliency detection. *IEEE Trans. Image Process.*, 22(10): 3766–3778.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1): 2096–2030.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Adv. Neural Inform. Process. Syst.*

Han, J.; Cheng, G.; Li, Z.; and Zhang, D. 2018. A Unified Metric Learning-Based Framework for Co-Saliency Detection. *IEEE Trans. Circ. Syst. Video Technol.*, 28(10): 2473–2483.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Conf. Comput. Vis. Pattern Recog.*, 9726–9735.

Hsu, K.-J.; Lin, Y.-Y.; and Chuang, Y.-Y. 2019. Deepco3: Deep instance co-segmentation by co-peak search and co-saliency detection. In *Conf. Comput. Vis. Pattern Recog.*, 8846–8855.

Jiang, H.; Yuan, Z.; Cheng, M.-M.; Gong, Y.; Zheng, N.; and Wang, J. 2013. Salient Object Detection: A Discriminative Regional Feature Integration Approach. *Int. J. Comput. Vis.*, 123: 251–268.

Jin, W.-D.; Xu, J.; Cheng, M.-M.; Zhang, Y.; and Guo, W. 2020. Icnet: Intra-saliency correlation network for co-saliency detection. In *Adv. Neural Inform. Process. Syst.*, 18749–18759.

Kim, J.; and Pavlovic, V. 2016. A Shape-Based Approach for Salient Object Detection Using Deep Learning. In *Eur. Conf. Comput. Vis.*, 455–470.

Li, B.; Sun, Z.; Tang, L.; Sun, Y.; and Shi, J. 2019. Detecting Robust Co-Saliency with Recurrent Co-Attention Neural Network. In *Int. Joint Conf. Artif. Intell.*, 818–825.

Li, G.; and Yu, Y. 2015. Visual saliency based on multiscale deep features. In *Conf. Comput. Vis. Pattern Recog.*, 5455–5463.

Li, G.; and Yu, Y. 2016. Deep Contrast Learning for Salient Object Detection. In *Conf. Comput. Vis. Pattern Recog.*, 478–487.

Li, X.; Lu, H.; Zhang, L.; Ruan, X.; and Yang, M.-H. 2013. Saliency Detection via Dense and Sparse Reconstruction. In *Int. Conf. Comput. Vis.*, 2976–2983.

Lin, T.-Y.; Dollár, P.; Girshick, R. B.; He, K.; Hariharan, B.; and Belongie, S. J. 2017. Feature Pyramid Networks for Object Detection. In *Conf. Comput. Vis. Pattern Recog.*, 936–944.

Liu, J.-J.; Hou, Q.; Liu, Z.-A.; and Cheng, M.-M. 2022. PoolNet+: Exploring the Potential of Pooling for Salient Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1–1.

Liu, N.; Han, J.; and Yang, M.-H. 2018. PiCANet: Learning Pixel-Wise Contextual Attention for Saliency Detection. In *Conf. Comput. Vis. Pattern Recog.*, 3089–3098.

Liu, Y.; Zhang, Q.; Zhang, D.; and Han, J. 2019. Employing Deep Part-Object Relationships for Salient Object Detection. In *Int. Conf. Comput. Vis.*, 1232–1241.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *Int. Conf. Learn. Represent.*

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inform. Process. Syst.*, volume 32.

Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; and Jagersand, M. 2019. Basnet: Boundary-aware salient object detection. In *Conf. Comput. Vis. Pattern Recog.*, 7479–7489.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39: 1137–1149.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Med. Image. Comput. Comput. Assist. Interv.*, 234–241.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Conf. Comput. Vis. Pattern Recog.*, 815–823.

Su, Y.; Deng, J.; Sun, R.; Lin, G.; and Wu, Q. 2022. A Unified Transformer Framework for Group-based Segmentation: Co-Segmentation, Co-Saliency Detection and Video Salient Object Detection. *arXiv preprint arXiv:2203.04708*.

Tang, L. 2021. CoSformer: Detecting Co-Salient Object with Transformers. *arXiv preprint arXiv:2104.14729*.

Tang, L.; Li, B.; Kuang, S.; Song, M.; and Ding, S. 2022. Re-thinking The Relations in Co-saliency Detection. *IEEE Trans. Circ. Syst. Video Technol.*, 1–1.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022. Pvtv2: Improved baselines with pyramid vision transformer. *Comput. Vis. Media*, 8(3): 1–10.

Wang, C.; Zha, Z.; Liu, D.; and Xie, H. 2019. Robust Deep Co-Saliency Detection with Group Semantic. In *AAAI Conf. Art. Intell.*, 8917–8924.

Wang, L.; Lu, H.; Ruan, X.; and Yang, M.-H. 2015. Deep networks for saliency detection via local estimation and global search. In *Conf. Comput. Vis. Pattern Recog.*, 3183–3192.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In *Int. Conf. Comput. Vis.*, 548–558.

Wang, X.; Girshick, R. B.; Gupta, A. K.; and He, K. 2018. Non-local Neural Networks. In *Conf. Comput. Vis. Pattern Recog.*, 7794–7803.

Wang, X.; Jabri, A.; and Efros, A. A. 2019. Learning correspondence from the cycle-consistency of time. In *Conf. Comput. Vis. Pattern Recog.*, 2566–2576.

Wei, L.; Zhao, S.; Bourahla, O. E. F.; Li, X.; and Wu, F. 2017. Group-wise Deep Co-saliency Detection. In *Int. Joint Conf. Artif. Intell.*, 3041–3047.

Winn, J.; Criminisi, A.; and Minka, T. 2005. Object categorization by learned universal visual dictionary. In *Int. Conf. Comput. Vis.*, volume 2, 1800–1807.

Wu, Z.; Su, L.; and Huang, Q. 2019. Stacked Cross Refinement Network for Edge-Aware Salient Object Detection. In *Int. Conf. Comput. Vis.*, 7263–7272.

Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2017. Joint Detection and Identification Feature Learning for Person Search. In *Conf. Comput. Vis. Pattern Recog.*, 3376–3385.

Yu, S.; Xiao, J.; Zhang, B.; and Lim, E. G. 2022. Democracy Does Matter: Comprehensive Feature Mining for Co-Salient Object Detection. In *Conf. Comput. Vis. Pattern Recog.*

Zeng, Y.; Zhuge, Y.; Lu, H.; and Zhang, L. 2019. Joint learning of saliency detection and weakly supervised semantic segmentation. In *Int. Conf. Comput. Vis.*, 7223–7233.

Zhang, D.; Han, J.; Li, C.; Wang, J.; and Li, X. 2016a. Detection of co-salient objects by looking deep and wide. *Int. J. Comput. Vis.*, 120(2): 215–232.

Zhang, D.; Meng, D.; and Han, J. 2017. Co-Saliency Detection via a Self-Paced Multiple-Instance Learning Framework. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(5): 865–878.

Zhang, H.; Zhang, H.; Wang, C.; and Xie, J. 2019. Co-Occurrent Features in Semantic Segmentation. In *Conf. Comput. Vis. Pattern Recog.*, 548–557.

Zhang, J.; Sclaroff, S.; Lin, Z. L.; Shen, X.; Price, B. L.; and Mech, R. 2016b. Unconstrained Salient Object Detection via Proposal Subset Optimization. In *Conf. Comput. Vis. Pattern Recog.*, 5733–5742.

Zhang, K.; Dong, M.; Liu, B.; Yuan, X.-T.; and Liu, Q. 2021. DeepACG: Co-Saliency Detection via Semantic-aware Contrast Gromov-Wasserstein Distance. In *Conf. Comput. Vis. Pattern Recog.*, 13698–13707.

Zhang, K.; Li, T.; Shen, S.; Liu, B.; Chen, J.; and Liu, Q. 2020a. Adaptive Graph Convolutional Network With Attention Graph Clustering for Co-Saliency Detection. In *Conf. Comput. Vis. Pattern Recog.*, 9047–9056.

Zhang, Q.; Cong, R.; Hou, J.; Li, C.; and Zhao, Y. 2020b. CoADNet: Collaborative aggregation-and-distribution networks for co-salient object detection. In *Adv. Neural Inform. Process. Syst.*, 6959–6970.

Zhang, X.; Wang, T.; Qi, J.; Lu, H.; and Wang, G. 2018. Progressive Attention Guided Recurrent Network for Salient Object Detection. In *Conf. Comput. Vis. Pattern Recog.*, 714–722.

Zhang, Z.; Jin, W.; Xu, J.; and Cheng, M.-M. 2020c. Gradient-induced co-saliency detection. In *Eur. Conf. Comput. Vis.*, 455–472.

Zhao, J.-X.; Liu, J.; Fan, D.-P.; Cao, Y.; Yang, J.; and Cheng, M.-M. 2019. EGNet: Edge Guidance Network for Salient Object Detection. In *Int. Conf. Comput. Vis.*, 8778–8787.

Zhao, R.; Ouyang, W.; Li, H.; and Wang, X. 2015. Saliency detection by multi-context deep learning. In *Conf. Comput. Vis. Pattern Recog.*, 1265–1274.

Zhao, T.; and Wu, X. 2019. Pyramid Feature Attention Network for Saliency Detection. In *Conf. Comput. Vis. Pattern Recog.*, 3080–3089.

Zheng, P.; Fu, H.; Fan, D.-P.; Fan, Q.; Qin, J.; and Van Gool, L. 2022. GCoNet+: A Stronger Group Collaborative Co-Salient Object Detector. *arXiv preprint arXiv:2205.15469*.

Zhou, Z.; Pei, W.; Li, X.; Wang, H.; Zheng, F.; and He, Z. 2021. Saliency-Associated Object Tracking. In *Conf. Comput. Vis. Pattern Recog.*, 9846–9855.

Zhuge, M.; Fan, D.-P.; Liu, N.; Zhang, D.; Xu, D.; and Shao, L. 2022. Salient Object Detection via Integrity Learning. *IEEE Trans. Pattern Anal. Mach. Intell.*