Co-Salient Object Detection with Uncertainty-aware Group Exchange-Masking

Yang Wu[1] Huihui Song[1*] Bo Liu[2] Kaihua Zhang[1] Dong Liu[3]
[1]B-DAT and CICAEET, Nanjing University of Information Science and Technology, Nanjing, China
[2]Walmart Global Tech, Sunnyvale, CA, 94086, USA
[3]Netflix Inc, Los Gatos, CA, 95032, USA
songhuihui@nuist.edu.cn

Abstract

The traditional definition of co-salient object detection (CoSOD) task is to segment the common salient objects in a group of relevant images. Existing CoSOD models by-default adopt the group consensus assumption. This brings about model robustness defect under the condition of irrelevant images in the testing image group, which hinders the use of CoSOD models in real-world applications. To address this issue, this paper presents a group exchange-masking (GEM) strategy for robust CoSOD model learning. With two group of image containing different types of salient object as input, the GEM first selects a set of images from each group by the proposed learning based strategy, then these images are exchanged. The proposed feature extraction module considers both the uncertainty caused by the irrelevant images and group consensus in the remaining relevant images. We design a latent variable generator branch which is made of conditional variational autoencoder to generate uncertainly-based global stochastic features. A CoSOD transformer branch is devised to capture the correlation-based local features that contain the group consistency information. At last, the output of two branches are concatenated and fed into a transformer-based decoder, producing robust co-saliency prediction. Extensive evaluations on co-saliency detection with and without irrelevant images demonstrate the superiority of our method over a variety of state-of-the-art methods.

1. Introduction

Co-salient object detection (CoSOD) is to segment the common salient objects in a group of relevant images. By detecting the co-salient object in a group of images, the images' background and redundant content are removed, which helps the downstream tasks such as object tracking [38], co-segmentation [48] and video co-localization [14], to name a few.

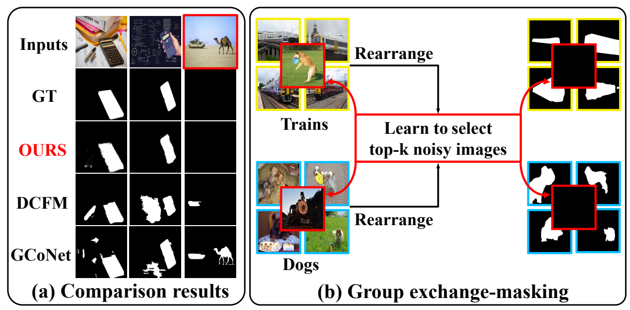

Figure 1. (a) When there exists an irrelevant image in the test group, the state-of-the-art CoSOD models such as DCFM [39] and GCoNet [9] tend to generate false positive predictions for the noisy image, yet our method can achieve an accurate prediction due to the use of group exchange-masking strategy in (b).

Group consensus assumption is widely used by existing CoSOD models, *i.e.*, these models presume that all images in the same group contain the common salient targets. The widely-used CoSOD benchmark datasets, such as COCO-SEG [33], CoCA [50], and CoSOD3k [8] organize the training and testing images that contain the same object as a group. Many existing CoSOD models consider the group consensus characteristic in modeling. For example, early works such as [2, 10, 13] extract hand-crafted features for inter-image co-object correspondence discovery. The deep learning models proposed in [15, 20, 36, 39, 44, 46, 48] use one group of relevant images as training data input for consensus representation learning. Among them, a variety of novel model design techniques have been developed to make full use of the group consistency characteristic, such as the low-rank feature learning [46], co-attention model [20] and intra-saliency correlation learning [15]. The issues of the group consensus assumption are partially studied in recent literature [9], which further models the inter-

*Corresponding author. This work is supported in part by National Key Research and Development Program of China under Grant No. 2018AAA0100400, in part by the NSFC under Grant Nos. 62276141, 61872189.

group separability for more discriminative feature learning by a group collaborating module.

In this paper, we find that the group consensus assumption also restricts the CoSOD model's robustness against the images without common object. As illustrated by Figure 1(a), state-of-the-art CoSOD models tend to output false positive predictions for the irrelevant image. This issue hinders the use of CoSOD models in real-world applications where the testing inputs are likely to contain irrelevant images. To enhance the model's robustness, we propose a learning framework called group exchange-masking (GEM). The GEM is illustrated by Figure 1(b). Given two image groups that contain different types of co-salient objects, we exchange several images between one group and the other. Those exchanged images are called noisy images. The number of noisy images is chosen to be less than the number of remaining relevant images in the group, so that the co-salient object in the noisy images forms a negative object but not the dominant co-salient object. The "masking" strategy refers to the label regeneration of the noisy images. Because there is no co-salient object, in the regenerated label, the original ground-truth object is masked. The learning objective is to correctly predict both the co-salient objects in the original relevant images and the added noisy images.

Adding noisy images to the training image group brings about uncertainty to the CoSOD model learning since there is some probability of no expected common object in each image. We design a dual-path image feature extraction module to model the group uncertainty in addition to the group consensus property. Specifically, we design a latent variable generator branch (LVGB) to extract the uncertainty-based global image features. The LVGB module is motivated by the conditional variational autoencoder (CVAE) [32] that is widely used to address the uncertainty in vision applications including image background modeling [19], RGB-D saliency detection [43] and image reconstruction [41]. In parallel with LVGB, we feed the image group into a CoSOD Transformer Branch (CoSOD-TB). The CoSOD-TB partitions each image group into local patches, and the attention mechanism in the transformer enables this branch to model patch-wise correlation-based local features. As a result, the group consistency information can be captured by this branch. The outputs of the two branches are concatenated and fed into a transformer-based decoder for co-saliency prediction. The proposed model has the following technical contributions.

- A robust CoSOD model learning mechanism, called group exchange-masking is proposed. By exchanging images between two groups, we augment the training data containing irrelevant images as noise to enhance model's robustness. This is different from the traditional CoSOD model learning frameworks that use groups of relevant images as training data.
- We propose a dual-path feature extraction module composed of the LVGB and the CoSOD-TB. The LVGB is designed to model the uncertainty of co-salient object existence. The CoSOD-TB is for the consensus feature extraction of the salient object in the relevant images.
- Extensive evaluations on three benchmark datasets, including CoSal2015 [42], CoCA [15], and CoSOD3k [7] show that the superiority of the proposed model to the state-of-the-art methods in terms of all evaluation metrics. Besides, the proposed model demonstrates good robustness for dealing with noisy data without co-salient objects.

2. Related Work

2.1. Co-salient Object Detection

In the past, CoSOD methods used to extract handcrafted features such as Gabor and SIFT features from images and then detect co-saliency by utilizing the consistency of low-level features between the images being tested [2]. A series of studies attempt to capture the intra-image constraints by employing a manifold ranking scheme to produce saliency maps [21], or using a global association constraint with clustering [10], or translational alignment [13]. More recently, there is a surge of deep learning-based CoSOD models that learn feature representation and saliency predictor in an end-to-end manner [9, 12, 36, 46, 47, 50]. Wei *et al.* [36] designs a collaborative learning framework for CoSOD that discovers the collaborative and interactive relationships between intra-image and single-image feature representations in a group-wise manner. Hsu *et al.* [12] present an unsupervised CNN-based model for CoSOD. In [46], a hierarchical framework is proposed for CoSOD where the initial CoSOD results generated by the CNN model is refined by label smoothing. Zhang *et al.* [50] propose a gradient-induced model for CoSOD that utilizes the image gradient information to induce more attentions to the discriminative co-salient feature learning. In [47], a deep graph neural network model is proposed to characterize the intra-image and inter-image region correspondence for CoSOD. In [9], a group collaborative learning strategy is proposed to explore inter-group relations for discriminating feature learning.

2.2. Robust Model and Feature Learning

In past years, with the development of deep learning, robust model and feature learning have received more and more attention as a possible solution to overcoming the bottleneck [28]. Various methods have been proposed to improve the robustness of the model. In [3] and [30], ways to

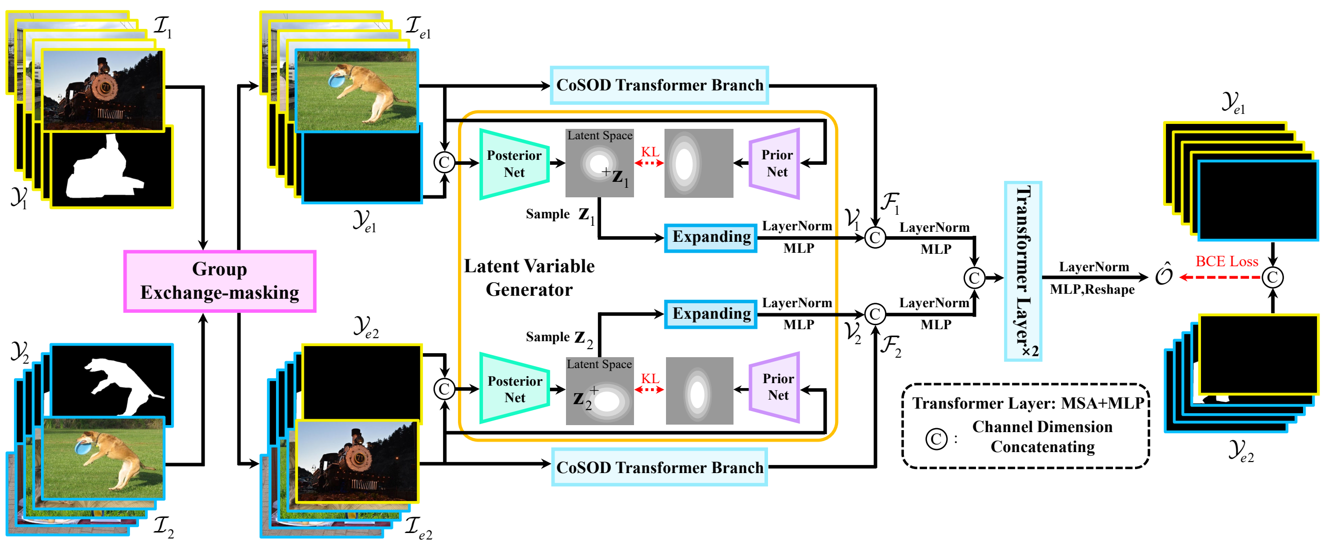

Figure 2. Pipeline of the proposed method. First, two groups of images $\{\mathcal{I}_1, \mathcal{I}_2\}$ with their masks $\{\mathcal{Y}_1, \mathcal{Y}_2\}$ are processed by group exchange-masking, yielding the exchanged images $\{\mathcal{I}_{e1}, \mathcal{I}_{e2}\}$ and their corresponding masked labels $\{\mathcal{Y}_{e1}, \mathcal{Y}_{e2}\}$. Then, $\{\mathcal{I}_{e1}, \mathcal{I}_{e2}\}$ and $\{\mathcal{Y}_{e1}, \mathcal{Y}_{e2}\}$ are passed through the latent variable generator branch, extracting the uncertainty-based global image features $\{\mathcal{V}_1, \mathcal{V}_2\}$ to eliminate the bias caused by noisy images. Meanwhile, $\{\mathcal{I}_{e1}, \mathcal{I}_{e2}\}$ are passed through the weight-shared CoSOD transformer branch, producing two groups of feature sequences $\{\mathcal{F}_1, \mathcal{F}_2\}$ with group consistency information and long-range independent information. Finally, $\mathcal{F}$ and $\mathcal{V}$ are concatenated and fed into the transformer decoder, yielding the predicted co-saliency maps $\hat{\mathcal{O}}$.

transfer the policy learned in a simulator to the real world are proposed to make features more robust. Rajeswaran *et al.* [29] learn a robust policy by sampling the worst case trajectories from a class of parametrized models, to learn a robust policy. Xie *et al.* [37] enhance the robustness of the model through randomly resizing and padding the training images. Several works leverage randomness to improve the robustness of models. Liu *et al.* [24] propose a noise layer that introduces randomness to both the input and the hidden layer output. Liu *et al.* [25] introduce a new min-max formulation that combines adversarial training with Bayesian Neural Networks, which achieves promising results.

3. Proposed Method

Figure 2 illustrates the pipeline of our framework. Given two groups of $2N$ images $\mathcal{I}_i = \{I_i^n \in \mathbb{R}^{H\times W\times 3}\}_{n=1}^N, i = 1, 2$ with height H and width W and manually-labeled binary masks $\mathcal{Y}_i = \{\mathbf{Y}_i^n \in \mathbb{R}^{H\times W}\}_{n=1}^N$ with different categories as input, we employ the GEM model to reorganize them, yielding $\mathcal{I}_{ei}$ and $\mathcal{Y}_{ei}$. Each pair of $\{\mathcal{I}_{ei}, \mathcal{Y}_{ei}\}$ are fed to the LVGB. In LVGB, we draw $\mathbf{z}_i \in \mathbb{R}^K$ from the Gaussian distribution $P_{\boldsymbol{\theta}}(\mathbf{z}_i|\mathcal{I}_{ei})$ with the learning parameters $\boldsymbol{\theta}$ and the posterior of $\mathbf{z}_i$ is formulated as $Q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathcal{I}_{ei}, \mathcal{Y}_{ei})$ with the learning parameters $\boldsymbol{\phi}$. We use the KL-Divergence metric to reduce the distance between $P_{\boldsymbol{\theta}}(\mathbf{z}_i|\mathcal{I}_{ei})$ and $Q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathcal{I}_{ei}, \mathcal{Y}_{ei})$. We further re-parameterize and expand $\mathbf{z}_i$ to yield a latent sequence $\mathcal{V}_i$ which has the same dimension as the feature sequences produced by transformers. Meanwhile, the image groups are passed through the CoSOD-TB with shared weights, producing the token sequences $\mathcal{F}_i$ which capture group consensus and long-range dependency information. At last, $\mathcal{V}_i$ and $\mathcal{F}_i$ are concatenated in channel dimension and passed through several subsequent processes as shown in Figure 2 followed by an up-sampling layer, yielding the corresponding predicted co-saliency maps $\hat{\mathcal{O}} = \{\hat{\mathbf{O}}^n \in \mathbb{R}^{H\times W}\}_{n=1}^{2N}$.

Algorithm 1 Group exchange-masking

Input: $\mathcal{I}_1 = \{I_1^n\}_{n=1}^N, \mathcal{Y}_1 = \{\mathbf{Y}_1^n\}_{n=1}^N$,
$\mathcal{I}_2 = \{I_2^n\}_{n=1}^N, \mathcal{Y}_2 = \{\mathbf{Y}_2^n\}_{n=1}^N$.
Output: $\mathcal{I}_{e1}, \mathcal{Y}_{e1}, \mathcal{I}_{e2}, \mathcal{Y}_{e2}$.
$\mathcal{I}_{e1}$, $\mathcal{Y}_{e1}$, $\mathcal{I}_{e2}$, $\mathcal{Y}_{e2}$ ← Rearranging $\mathcal{I}_1$, $\mathcal{Y}_1$, $\mathcal{I}_2$, $\mathcal{Y}_2$ via solving (1).
for $n = 1, 2, \ldots, k$ **do**
  $I_{e1}^n \leftarrow I_2^n, I_{e2}^n \leftarrow I_1^n$,
  $\mathbf{Y}_{e1}^n \leftarrow \mathbf{0} \in \mathbb{R}^{H\times W\times 1}$,
  $\mathbf{Y}_{e2}^n \leftarrow \mathbf{0} \in \mathbb{R}^{H\times W\times 1}$.
end for

3.1. Group Exchange-Masking

The widely-used group consensus assumption limits the robustness of the CoSOD model, especially when there are test images without common objects in the group. To this end, we reorganize the two input groups with the GEM tailored to CoSOD. The GEM is summarised in Algorithm 1:

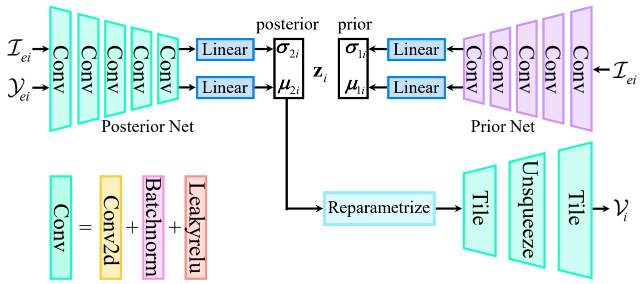

Figure 3. Architecture of the LVGB.

taking $\mathcal{I}_i$ and $\mathcal{Y}_i$ as input, the goal of GEM is to obtain $\mathcal{I}_{ei}$ and $\mathcal{Y}_{ei}$ which contain noisy images in each group and corresponding full masked ground truths. Specifically, we rearrange the two groups of images and their binary ground-truth masks in a decreasing order by solving (1). After this, we sample the top-k images of each group as noisy images and exchange them in each other's group, ensuring the maximum destructiveness of noisy images and the effectiveness of training.

This training process is essentially a min-max optimization process [26] as

$$\min_{\varphi} \sum_{n=1}^{k} \sum_{i=1}^{2} \max_{\mathcal{I}_i(n,:)} \mathcal{L}(f_{\varphi}(\mathcal{I}_i(n,:)), \hat{\mathcal{O}}_i(n,:)), \tag{1}$$

where the loss $\mathcal{L}$ is defined by (7), and f denotes our whole model with learnable parameters φ. Solving (1) finds out the most noisy images which can maximize the training loss, and our goal is to minimize the loss function w.r.t the intra-group noisy images to improve the model's robustness. At last, we mask the corresponding labels of the noisy images with the all-zeros maps.

3.2. Latent Variable Generator

By training the model with noisy images, the GEM is able to increase the model's robustness which can identify whether there exist co-objects or not. However, the model will focus on non-co-salient regions as a result of the uncertainty introduced by adding noisy images to the training image groups. This means that the model will become overconfident in the background regions in the early stage of training, leading to inaccurate prediction.

To address this issue, we further propose the LVGB that generates a low-dimensional latent space to represent the most common patterns in image groups [17]. The latent variables sampled from the feature space characterize uncertainties of group consensus and can be used to modulate the intermediate features from other branches to highlight the co-objects.

Figure 3 shows the details of our LVGB. The generator takes $\mathcal{I}_{ei}$ and latent variable $\mathbf{z}_i$ as input, and uses an encoder to produce stochastic prediction $P_{\boldsymbol{\omega}}(\mathcal{Y}_{ei}|\mathcal{I}_{ei}, \mathbf{z}_i)$ with learnable parameters $\boldsymbol{\omega}$, where $\mathcal{I}_{ei}$ is conditional variable, and $\mathcal{Y}_{ei}$ is output variable. $\mathbf{z}_i$ is drawn from the Gaussian distribution $P_{\boldsymbol{\theta}}(\mathbf{z}_i|\mathcal{I}_{ei})$, and the posterior of $\mathbf{z}_i$ is formulated as $Q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathcal{I}_{ei}, \mathcal{Y}_{ei})$ via an encoder that maps the input variable to the latent space. The loss of the LVG is defined as [32,43]

$$\begin{aligned}\mathcal{L}_{LVG} = \sum_{i=1}^{2} & E_{\mathbf{z}\sim Q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathcal{I}_{ei},\mathcal{Y}_{ei})}[-\log P_{\boldsymbol{\omega}}(\mathcal{Y}_{ei}|\mathcal{I}_{ei}, \mathbf{z}_i)] \\ & +\mathrm{KL}(Q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathcal{I}_{ei}, \mathcal{Y}_{ei})||P_{\boldsymbol{\theta}}(\mathbf{z}_i|\mathcal{I}_{ei})),\end{aligned} \tag{2}$$

where the Kullback-Leibler Divergence metric KL is defined as

$$\mathrm{KL}(Q||P) = \sum Q(x) \log \frac{Q(x)}{P(x)}. \tag{3}$$

In (2), the prior net $P_{\boldsymbol{\theta}}(\mathbf{z}_i|\mathcal{I}_{ei})$ is defined as a Gaussian distribution that maps the image group $\mathcal{I}_{ei}$ to a low-dimensional latent space that encodes the most common patterns in each group. We randomly sample prior latent variable $\mathbf{z}_i \sim \mathcal{N}(\boldsymbol{\mu}_{1i}, \mathrm{diag}(\boldsymbol{\sigma}_{1i}^2))$, where $\boldsymbol{\mu}_{1i}, \boldsymbol{\sigma}_{1i}^2 \in \mathbb{R}^K$ represent the mean and standard deviation vectors and $\boldsymbol{\theta}$ is the learnable parameter of the mapping function which consists five conv layers [43] as shown in Figure 3. In the posterior net, we use the similar encoder as the prior net to model $Q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathcal{I}_{ei}, \mathcal{Y}_{ei})$ that maps the concatenation of $\mathcal{I}_{ei}$ and $\mathcal{Y}_{ei}$ to the posterior latent variable $\mathbf{z}_i \sim \mathcal{N}(\boldsymbol{\mu}_{2i}, \mathrm{diag}(\boldsymbol{\sigma}_{2i}^2))$, where $\boldsymbol{\mu}_{2i}, \boldsymbol{\sigma}_{2i}^2 \in \mathbb{R}^K$. The KL in (2) is used to measure the difference between probability distributions prior $P_{\boldsymbol{\theta}}(\mathbf{z}_i|\mathcal{I}_{ei})$ and posterior $Q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathcal{I}_{ei}, \mathcal{Y}_{ei})$. It represents the information loss that occurs when using $Q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathcal{I}_{ei}, \mathcal{Y}_{ei})$ to approximate $P_{\boldsymbol{\theta}}(\mathbf{z}_i|\mathcal{I}_{ei})$, and a smaller value indicates greater similarity between the two distributions.

Each position in latent space symbolizes potential labeling changes or other potential factors that could lead to a variety of co-saliency predictions [17, 43]. The proposed GEM can meet the needs for diverse ground truths as we learn to select the most difficult samples to mask their ground truths to all-zeros maps. We hope diverse annotations in the posterior net $Q_{\boldsymbol{\phi}}(\mathbf{z}_i|\mathcal{I}_{ei}, \mathcal{Y}_{ei})$ during training can compel the prior net $P_{\boldsymbol{\theta}}(\mathbf{z}_i|\mathcal{I}_{ei})$ to encode labeling variants of the supplied inputs $\mathcal{I}_{ei}$. The statistics $\mathbf{z}_i$ is further processed by feature expanding [43]. Given a pair of $(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k)$ in each position of K dimensional vector, we parameterize them to obtain the latent vector

$$\mathbf{z}_k = \boldsymbol{\sigma}_k \odot \boldsymbol{\epsilon} + \boldsymbol{\mu}_k, \tag{4}$$

where $\boldsymbol{\epsilon}$ follows the standard normal distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{1})$. To fuse with the features $\mathcal{F}_i$ from the CoSOD-TB, as shown in Figure 3, we expand $\mathbf{z}_k$ to the feature sequence which has the same size as $\mathcal{F}_i$, yielding the stochastic feature $\mathcal{V}_i$. The visualization effect of $\mathcal{V}_i$ is partly shown in Fig-

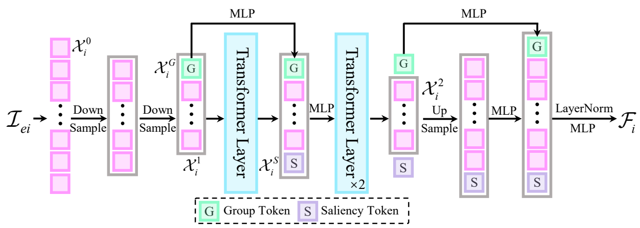

Figure 4. Architecture of the CoSOD-TB.

ure 5, as we can see, these features encode rich group consistency information that can well highlight the co-salient regions.

3.3. CoSOD Transformer

After getting the stochastic features $\mathcal{V}_i$ from the LVGB, we leverage these features to guide the learning of generating the general features $\mathcal{F}_i$ from the CoSOD-TB to focus on the co-salient regions (see the bottom row of Figure 5).

Figure 4 shows the architecture of the proposed CoSOD-TB. Each input image $I_i^n \in \mathcal{I}_{ei}$ is cropped into d patches with size $H/4 \times W/4$, and then all the image patches in each group construct the token sequences $\mathcal{X}_i^0 \in \mathbb{R}^{N \times \frac{H}{4} \times \frac{W}{4} \times 3d}$. Then, $\mathcal{X}_i^0$ is sent into the backbone of the transformer with a T2T architecture [40], obtaining the token sequences $\mathcal{X}_i^1 \in \mathbb{R}^{N \times \frac{H}{16} \times \frac{W}{16} \times c}$ which encode both local and global information. After this, we design the group token $\mathcal{X}^G \in \mathbb{R}^{N \times 1 \times c}$ to capture group-wise common information that is essential to extract co-objects, and the saliency token $\mathcal{X}^S \in \mathbb{R}^{N \times 1 \times c}$ to capture the specific information that encodes more co-object structure details. Afterwards, $\mathcal{X}^G$ and $\mathcal{X}^1$ are concatenated in the channel dimension and fed into the Transformer Layer as the design in ViT [4] to get the new token sequence. We further use an multi-layer perceptron (MLP) layer to integrate global information into $\mathcal{X}^G$. We fuse the $\mathcal{X}^S$ to integrate saliency information into sequences and split the $\mathcal{X}^G$ and $\mathcal{X}^S$ to up-sample the token sequence and get the token sequence $\mathcal{X}^2$. After undergoing some processing as shown in Figure 4, $\mathcal{F}_i$ is then concatenated with the stochastic features $\mathcal{V}_i$ followed by a decoder that includes two Transformer Layers and an MLP layer. During decoding, due to the use of stochastic features, the model is more possible to overcome the prejudices formed in the early training. As shown in the third row of Figure 5, the CoSOD-TB can integrate $\mathcal{V}_i$ and $\mathcal{F}_i$ well, helping the model focus on more reliable co-salient object regions. Finally, the output sequences are reshaped to produce the predicted co-saliency maps $\hat{\mathcal{O}}$.

The loss function of the CoSOD-TB is defined as

$$\mathcal{L}_{TRANS} = \frac{1}{2N} \sum_{n=1}^{N} \sum_{i=1}^{2} \ell_{BCE}(\mathcal{Y}_{ei}(n,:), \hat{\mathcal{O}}_i(n,:)), \quad (5)$$

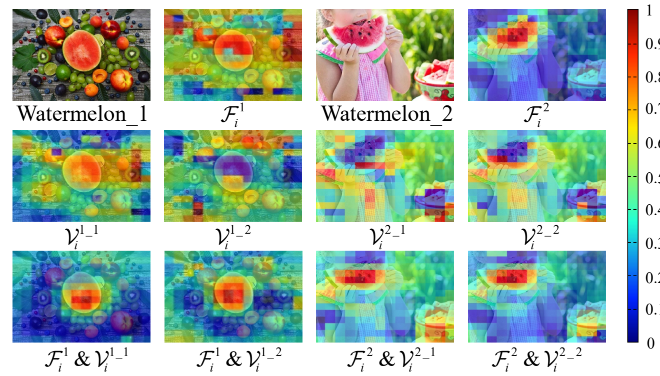

Figure 5. Comparison of effects of the "watermelon" group before and after fusing stochastic features. The first row: visualizations of the images and features obtained by the CoSOD-TB; The second row: visualizations of the features generated by the LVGB; The third row: visualizations of the integrated features.

where ℓ_{BCE} is a binary cross-entropy (BCE) loss [27] defined as

$$\begin{aligned}\ell_{BCE}(\mathcal{Y}(n,:),\hat{\mathcal{O}}(n,:)) = &-(\mathcal{Y}(n,:)^\top \log(\hat{\mathcal{O}}(n,:)) \\ &- (1-\mathcal{Y}(n,:))^\top \log(1-\hat{\mathcal{O}}(n,:))).\end{aligned} \quad (6)$$

3.4. Loss Function

The LVG and the CoSOD-TB are jointly trained in an end-to-end manner by optimizing the following multi-task loss

$$\mathcal{L} = \lambda_1 \mathcal{L}_{LVG} + \lambda_2 \mathcal{L}_{TRANS}, \quad (7)$$

where λ_1, λ_2 are the hyperparameters to balance each loss.

4. Experimental Results

4.1. Implementation Details

Our model is implemented under the PyTorch1.9.0 framework [27]. Our computing platform's acceleration is provided by a GeForce GTX 2080Ti GPU. The transformer backbone is the pre-trained T2T-ViT$_t$-14 [40] model since it has a similar computational complexity as CNNs-based ResNet50 [11] which is smaller than the VGG-16 [31] widely used in CoSOD. The training dataset are COCO-SEG released by [33] and DUTS released by [34]. They totally include about 208,250 images from 369 categories, and the relevant binary ground-truth masks are provided.

In each training episode, we randomly select two groups, which are processed by the proposed GEM with $k = 1$. Each group contains $N = 5$ images. The hyperparameters in (7) are set to $\lambda_1 = 0.25$ and $\lambda_2 = 1$. We resize the images to $224 \times 224 \times 3$ and use them as input. We train the network over 60,000 steps totally and the training process takes about 11 hours. During the training process, we use

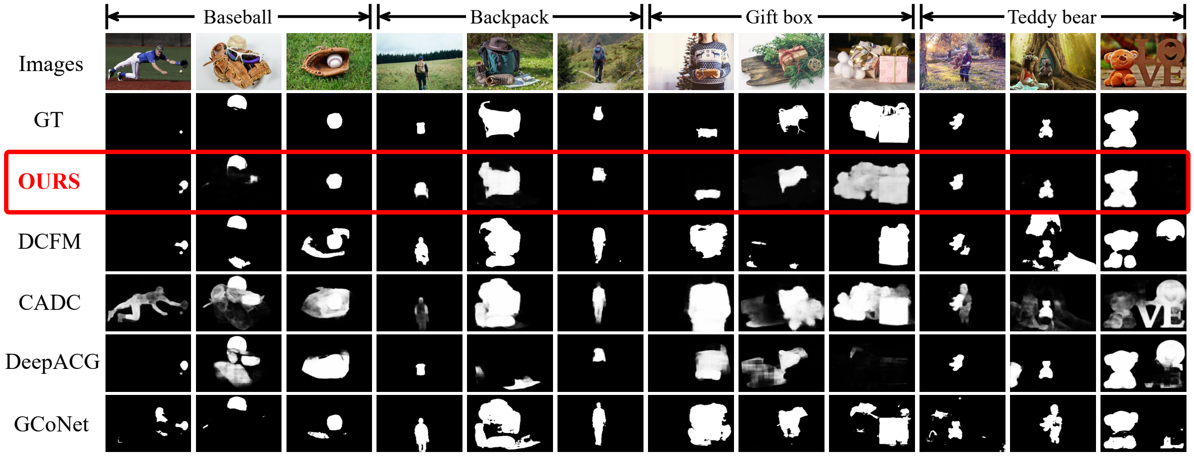

Figure 6. The qualitative comparisons with the recent state-of-the-art methods, including DCFM [39], CADC [49], DeepACG [45] and GCoNet [9].

the Adam algorithm [16] to optimize the whole network. The initial learning rate is set to $10e-4$ and decay to $10e-5$ at the 20,000-th epoch. The inference time is 31 *fps* in average, meeting the real-time application requirements.

4.2. Datasets and Evaluation Metrics

The test sets include CoSal2015 [42], CoCA [50] and CoSOD3k [8]. Among them, CoSal2015 contains $2,015$ images from 50 categories, and the objects therein have different sizes and shapes, which makes CoSal2015 challenging; CoCA is the newest dataset, including $1,297$ images from 80 categories. These images have extremely different styles and very complicated backgrounds, making it a very difficult data set; CoSOD3k contains $3,316$ images form 160 categories, which is by far the largest test set. It contains more disturbing targets, making it very challenging.

Four evaluation metrics are used for comparison, including MAE [33], E_{ϕ}^{max} [6], S_{α} [5] and F_{β}^{max} [1], which are to assess the average pixel-wise absolute difference, local and global similarity, structural similarity between the predictions and the ground truths, and the weighted harmonic mean of precision and recall, respectively.

4.3. Comparisons with State-of-the-art Methods

On the basis of unified evaluation codes [7] for fair comparison, we compare our method with several state-of-the-art methods published in recent three years, including RCAN [20], CSMG [46], SSNM [44], GCAGC [47], GICD [50], ICNet [15], GCoNet [9], DeepACG [45], CoEG-Net [7], HrSSMN [48], CADC [49], DCFM [39]. Among them, CADC [49] and DCFM [39] are the latest cutting-edge methods. **More experimental results can be found in the supplementary materials.**

Qualitative Results. Figure 6 shows some visual comparison results with four latest state-of-the-art methods, including DCFM, CADC, DeepACG and GCoNet. The selected four groups are very challenging and suffer from highly confusing interference co-objects, extremely complex backgrounds, and drastic scale changes. Specifically, in the group “Baseball”, the co-salient object is very small and the interference objects have similar shapes. Because all other methods are trapped in overconfidence, obsessed with the wrong results, and unable to extricate themselves. They cannot accurately detect the tiny objects and the background is wrongly segmented. In the group “Backpack”, benefiting from training with noisy samples, our model is more robust that can accurately locate co-salient objects and segment target details, while other methods DCFM, CADC and GCoNet misjudge the co-salient objects. Camouflage co-objects in complex backgrounds are generally acknowledged as difficult samples [18], and the co-salient items in the “Gift box” group are perfectly in place. Thanks to CoSOD-TB’s ability to capture long-range dependencies, our model learns strong feature representation that is able to accurately segment the camouflage co-salient objects, while CADC and GCoNet cannot finely segment the objects, and the other methods even fail in detecting the targets. The group “teddy bear” is also very challenging, and the co-salient objects in it suffer from interference objects and backgrounds with similar colors. Besides, the objects also have very different sizes. In this case, our model still achieves promising results, demonstrating its strong robustness to a variety of challenging factors. However, other methods yield a large number of inaccurate segmentation regions.

Quantitative Results. Figure 7 shows the PR and F-measure curves of our method and other state-of-the-art

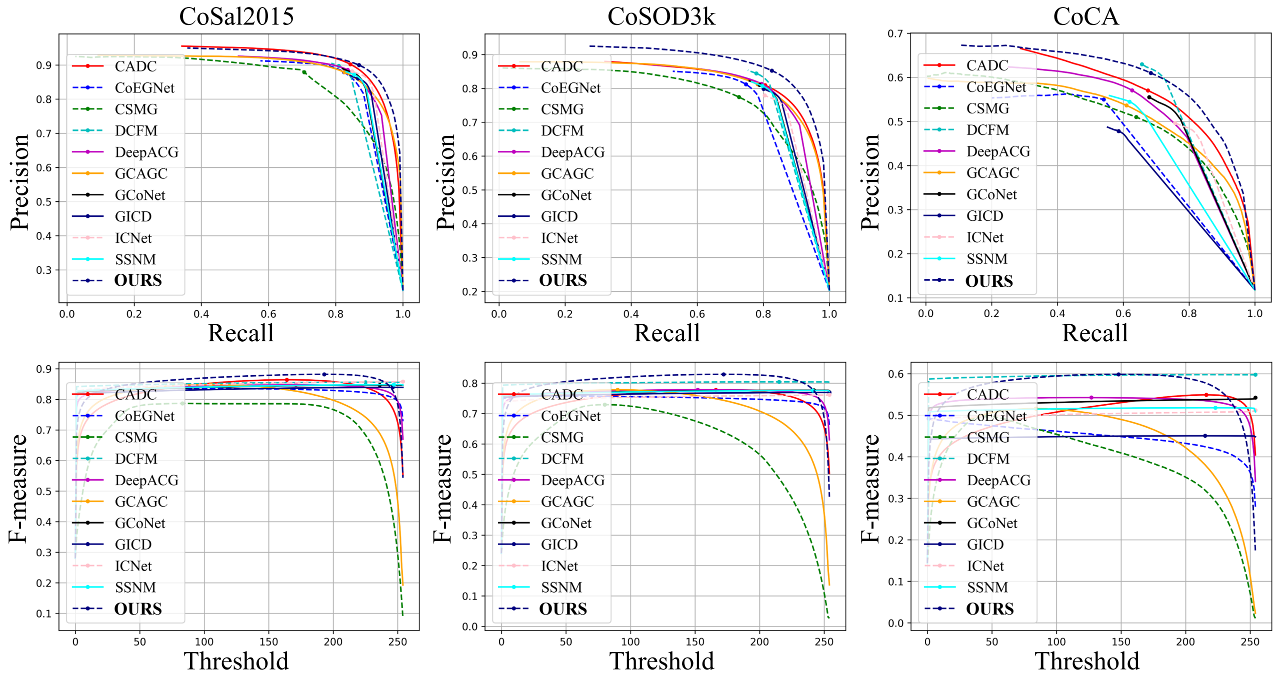

Figure 7. Comparisons with state-of-the-art methods since the year of 2019 in terms of PR and F-measure curves on three benchmark datasets.

Table 1. Statistic comparisons of our model with the other state-of-the-arts since the year of 2019.

Methods	CoSal2015				CoSOD3k				CoCA			
	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{max}\uparrow$	$F_\beta^{max}\uparrow$	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{max}\uparrow$	$F_\beta^{max}\uparrow$	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{max}\uparrow$	$F_\beta^{max}\uparrow$
RCAN(IJCAI2019)	0.126	0.779	0.842	0.764	0.130	0.744	0.808	0.688	0.160	0.616	0.702	0.422
CSMG(CVPR2019)	0.130	0.774	0.818	0.777	0.157	0.711	0.723	0.645	0.124	0.632	0.734	0.503
SSNM(AAAI2020)	0.102	0.788	0.843	0.794	0.120	0.726	0.756	0.675	0.116	0.628	0.741	0.482
GCAGC(CVPR2020)	0.085	0.817	0.866	0.813	0.100	0.785	0.816	0.740	0.118	0.669	0.754	0.523
GICD(ECCV2020)	0.072	0.842	0.884	0.834	0.089	0.794	0.831	0.743	0.125	0.658	0.701	0.504
ICNet(NIPS2020)	0.058	0.857	0.900	0.858	0.089	0.794	0.845	0.762	0.147	0.654	0.705	0.514
CoEGNet(TPAMI2021)	0.077	0.836	0.882	0.832	0.092	0.762	0.825	0.736	0.106	0.612	0.717	0.493
GCoNet(CVPR2021)	0.069	0.845	0.887	0.847	0.071	0.802	0.860	0.750	0.105	0.673	0.760	0.524
DeepACG(CVPR2021)	0.066	0.853	0.893	0.847	0.079	0.811	0.859	0.779	0.104	0.685	0.759	0.564
CADC(ICCV2021)	0.064	0.866	0.906	0.862	0.096	0.801	0.840	0.759	0.132	0.681	0.744	0.548
HrSSMN(TMM2022)	0.062	0.845	0.895	0.841	0.087	0.788	0.842	0.753	0.106	0.671	0.739	0.532
DCFM(CVPR2022)	0.067	0.838	0.892	0.856	0.067	0.810	0.874	0.805	**0.085**	0.710	0.783	0.598
OURS	**0.053**	**0.885**	**0.933**	**0.882**	**0.061**	**0.853**	**0.911**	**0.829**	0.095	**0.726**	**0.808**	**0.599**

methods. As we can see, the PR curves produced by our method wrap around the curves generated by other methods. Also, all the F-measure curves of our method are above those generated by other methods.

Meanwhile, we list the statistic comparison results of all compared methods in Table 1. It is clear that our approach achieves the most competitive result compared to its counterparts. Specifically, on the CoSal2015 dataset, our method achieves the best scores of 0.053, 0.885, 0.933, and 0.882 in terms of all metrics, with a significant gain of 0.5%, 1.9%, 2.7% and 2.0%, respectively, compared to the second-best preforming ICNet [15] and CADC [49]. On the most recent and difficult datasets, CoSOD3k and CoCA, our method also achieves the best performance. Especially on CoSOD3k, our method reaches the best scores of 0.061, 0.853, 0.911 and 0.829 in terms of MAE, E_ϕ^{max}, S_α and F_β^{max}, respectively, with a great gain of 0.6%, 4.3%, 3.7% and 2.4% compared to the second-best performing DCFM [39]. The effectiveness of our method has been fully validated on the large-scale and challenging datasets.

4.4. Ablation Study

To verify the effectiveness of our designs, we conduct the ablation study of our method on three datasets. We have made a lot of changes to VST [23] and regard the modified version as the baseline. Table 2 lists the results, and we can observe that every key design in our model makes a significant contribution to the overall performance. Taking

Table 2. Ablations of our method on the effectiveness of the GEM , LVGB, and CoSOD-TB.

Strategies			CoSal2015				CoSOD3k				CoCA			
GEM	LVGB	CoSOD-TB	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{max}\uparrow$	$F_\beta^{max}\uparrow$	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{max}\uparrow$	$F_\beta^{max}\uparrow$	$MAE\downarrow$	$S_\alpha\uparrow$	$E_\phi^{max}\uparrow$	$F_\beta^{max}\uparrow$
			0.060	0.854	0.887	0.860	0.075	0.787	0.863	0.778	0.105	0.710	0.785	0.564
✓			0.061	0.883	0.928	0.877	0.065	0.834	0.880	0.817	0.109	0.710	0.788	0.571
	✓		0.058	0.877	0.926	0.872	0.063	0.840	0.871	0.805	0.100	0.719	0.789	0.574
		✓	0.053	0.874	0.893	0.880	0.062	0.842	0.900	0.814	0.098	0.716	0.798	0.583
	✓	✓	0.055	0.869	0.930	0.876	0.063	0.847	0.906	0.819	0.104	0.724	0.802	0.597
✓	✓		0.060	0.886	0.925	0.872	0.069	0.850	0.895	0.823	0.100	0.718	0.792	0.587
✓		✓	0.054	0.880	0.919	0.878	0.061	0.849	0.889	0.829	0.096	0.720	0.805	0.595
✓	✓	✓	**0.053**	**0.885**	**0.933**	**0.882**	**0.061**	**0.853**	**0.911**	**0.829**	**0.095**	**0.726**	**0.808**	**0.599**

Table 3. Complexity analyses. The number of inputs is set to 5 to maintain unity.

methods	FLOPs(G)↓	param.(M)↓	runtime(fps)↑	$F_\beta^{max}\uparrow$
GICD(ECCV20)	364.7	278.0	40.8	0.504
GCoNet(CVPR21)	259.9	142.0	**116.2**	0.524
CADC(ICCV21)	330.0	392.8	18.0	0.548
DCFM(CVPR22)	251.9	142.3	84.4	0.598
OURS	**211.8**	**52.3**	31.0	**0.599**

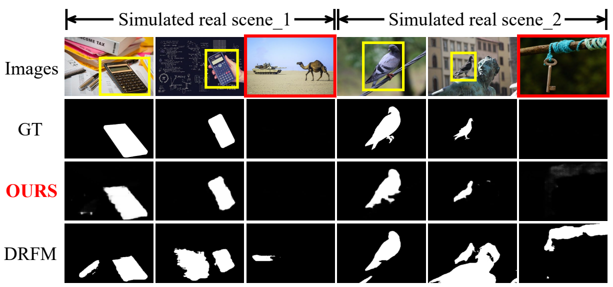

Figure 8. Comparison with the current most competitive method DCFM [39] in simulated real scenes.

the results in CoSal2015 as examples, without GEM, the performance of our model gets worse in terms of all metrics especially for the S_α that drops from 0.885 to 0.869 by 1.6%. With the GEM alone, the performance of our model improves greatly, and especially S_α and E_ϕ^{max} have a gain of 2.9% and 4.1%, respectively. Also, LVGB achieves the same excellent effect as GEM, because the uncertainty features provided by the LVGB can help the model consider more possibilities to enhance the robustness and overcome the over-confidence to non-salient regions. The CoSOD-TB makes important contributions to fine segmentation of images. With the CoSOD-TB, the MAE gets 0.7% better from 0.060 to 0.053. The results on CoSOD3k and CoCA show the same trend, proving the effectiveness of each design of our method.

4.5. Practical Application

For the practicability of the algorithm, we conduct complexity comparisons with the same settings as DCFM [39]. The results can be seen in Table 3. Our method achieves both smallest FLOPs, parameters and best performance in terms of F_β^{max}. Meanwhile, the runtime of our method is 31 *fps*, meeting the real-time requirements. This shows that our method is more suitable for deployment to various application devices.

We further test the performance of the model in the real scenes. In the real-life scenarios, there may not always exist co-salient objects within the group of images [22]. Figure 8 shows the images in the simulated real scenes that may not always exist co-objects. In these cases, the model should segment the co-objects existing in most images, while outputting all-zero masks on images without co-objects. Our method is better than the most competitive method DCFM [39] in dealing with real situations. When dealing with the images without co-objects, our model produces all-zero maps as much as possible. It shows that our method is more practical than other methods. **More experimental results can be found in the supplementary materials.**

5. Conclusion

In this paper, we have discovered that the paradigm of group consensus assumption has reduced the model's robustness and practical application value when confronted with irrelevant images in groups. A CoSOD model learning framework that is distinct from the classic CoSOD model learning framework has been developed. First, the group exchange-masking strategy has been devised, which is capable of automatically selecting the most informative noisy images from two groups and exchanging them in each other's groups to assist the model in learning more robust representations encoded with rich group consensus information. Second, the latent variable generator branch has been created to provide uncertainly-based global stochastic features that can regulate intermediate features from other branches to focus on co-objects. Third, the CoSOD transformer branch has been created to capture correlation-based global characteristics that carry information about group consistency. These branches' features are concatenated and put into a transformer-based decoder, yielding high-quality co-saliency maps. Extensive evaluations with and without irrelevant images have demonstrated the superiority of our method over a variety of state-of-the-art methods. In the future, we plan to build a new dataset for real scenarios that can provide a platform to evaluate model robustness and practical value in a more reliable way. Further more, we will consider the potential application of our method in defending adversarial attacks [35].

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1597–1604. IEEE, 2009. 6

[2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2129–2136. IEEE, 2011. 1, 2

[3] Paul Christiano, Zain Shah, Igor Mordatch, Jonas Schneider, Trevor Blackwell, Joshua Tobin, Pieter Abbeel, and Wojciech Zaremba. Transfer from simulation to real world through learning deep inverse dynamics model. *arXiv preprint arXiv:1610.03518*, 2016. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[5] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4548–4557, 2017. 6

[6] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018. 6

[7] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 6

[8] Deng-Ping Fan, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Huazhu Fu, and Ming-Ming Cheng. Taking a deeper look at co-salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2919–2929, 2020. 1, 6

[9] Qi Fan, Deng-Ping Fan, Huazhu Fu, Chi-Keung Tang, Ling Shao, and Yu-Wing Tai. Group collaborative learning for co-salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12288–12298, 2021. 1, 2, 6

[10] Huazhu Fu, Xiaochun Cao, and Zhuowen Tu. Cluster-based co-saliency detection. *IEEE Transactions on Image Processing*, 22(10):3766–3778, 2013. 1, 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[12] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Co-attention cnns for unsupervised object co-segmentation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 748–756, 2018. 2

[13] David E Jacobs, Dan B Goldman, and Eli Shechtman. Cosaliency: Where people look when comparing images. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, pages 219–228, 2010. 1, 2

[14] Koteswar Rao Jerripothula, Jianfei Cai, and Junsong Yuan. Cats: Co-saliency activated tracklet selection for video co-localization. In *Proceedings of the European Conference on Computer Vision*, pages 187–202. Springer, 2016. 1

[15] Wen-Da Jin, Jun Xu, Ming-Ming Cheng, Yi Zhang, and Wei Guo. Icnet: Intra-saliency correlation network for co-saliency detection. *Advances in Neural Information Processing Systems*, 33:18749–18759, 2020. 1, 2, 6, 7

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[17] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *Advances in Neural Information Processing Systems*, 31, 2018. 4

[18] Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10071–10081, 2021. 6

[19] Bo Li, Zhengxing Sun, and Yuqi Guo. Supervae: Superpixelwise variational autoencoder for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8569–8576, 2019. 2

[20] Bo Li, Zhengxing Sun, Lv Tang, Yunhan Sun, and Jinlong Shi. Detecting robust co-saliency with recurrent co-attention neural network. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 2, page 6, 2019. 1, 6

[21] Yijun Li, Keren Fu, Zhi Liu, and Jie Yang. Efficient saliency-model-guided visual co-saliency detection. *IEEE Signal Processing Letters*, 22(5):588–592, 2014. 2

[22] Jiawei Liu, Jing Zhang, Kaihao Zhang, and Nick Barnes. Generalised co-salient object detection. *arXiv preprint arXiv:2208.09668*, 2022. 8

[23] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 7

[24] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision*, pages 369–385, 2018. 3

[25] Xuanqing Liu, Yao Li, Chongruo Wu, and Cho-Jui Hsieh. Adv-bnn: Improved adversarial defense through robust bayesian neural network. *arXiv preprint arXiv:1810.01279*, 2018. 3

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 4

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019. 5

[28] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pages 2817–2826. PMLR, 2017. 2

[29] Aravind Rajeswaran, Sarvjeet Ghotra, Balaraman Ravindran, and Sergey Levine. Epopt: Learning robust neural network policies using model ensembles. *arXiv preprint arXiv:1610.01283*, 2016. 3

[30] Andrei A Rusu, Matej Večerík, Thomas Rothörl, Nicolas Heess, Razvan Pascanu, and Raia Hadsell. Sim-to-real robot learning from pixels with progressive nets. In *Conference on robot learning*, pages 262–270. PMLR, 2017. 2

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[32] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems*, 28, 2015. 2, 4

[33] Chong Wang, Zheng-Jun Zha, Dong Liu, and Hongtao Xie. Robust deep co-saliency detection with group semantic. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8917–8924, 2019. 1, 5, 6

[34] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 136–145, 2017. 5

[35] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019. 8

[36] Lina Wei, Shanshan Zhao, Omar El Farouk Bourahla, Xi Li, and Fei Wu. Group-wise deep co-saliency detection. *arXiv preprint arXiv:1707.07381*, 2017. 1, 2

[37] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017. 3

[38] Xi Yang, Shaoyi Li, Jun Ma, Jun-yan Yang, and Jie Yan. Co-saliency-regularized correlation filter for object tracking. *Signal Processing: Image Communication*, 103:116655, 2022. 1

[39] Siyue Yu, Jimin Xiao, Bingfeng Zhang, and Eng Gee Lim. Democracy does matter: Comprehensive feature mining for co-salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 979–988, 2022. 1, 6, 7, 8

[40] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 5

[41] Chen Zhang, Riccardo Barbano, and Bangti Jin. Conditional variational autoencoder for learned image reconstruction. *Computation*, 9(11):114, 2021. 2

[42] Dingwen Zhang, Junwei Han, Chao Li, and Jingdong Wang. Co-saliency detection via looking deep and wide. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2994–3002, 2015. 2, 6

[43] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Saleh, Sadegh Aliakbarian, and Nick Barnes. Uncertainty inspired rgb-d saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 4

[44] Kaihua Zhang, Jin Chen, Bo Liu, and Qingshan Liu. Deep object co-segmentation via spatial-semantic network modulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12813–12820, 2020. 1, 6

[45] Kaihua Zhang, Mingliang Dong, Bo Liu, Xiao-Tong Yuan, and Qingshan Liu. Deepacg: Co-saliency detection via semantic-aware contrast gromov-wasserstein distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13703–13712, 2021. 6

[46] Kaihua Zhang, Tengpeng Li, Bo Liu, and Qingshan Liu. Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3095–3104, 2019. 1, 2, 6

[47] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9050–9059, 2020. 2, 6

[48] Kaihua Zhang, Yang Wu, Mingliang Dong, Bo Liu, Dong Liu, and Qingshan Liu. Deep object co-segmentation and co-saliency detection via high-order spatial-semantic network modulation. *IEEE Transactions on Multimedia*, 2022. 1, 6

[49] Ni Zhang, Junwei Han, Nian Liu, and Ling Shao. Summarize and search: Learning consensus-aware dynamic convolution for co-saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4167–4176, 2021. 6, 7

[50] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In *Proceedings of the European Conference on Computer Vision*, pages 455–472. Springer, 2020. 1, 2, 6