

Generalizable Co-Salient Object Detection via Mixed Content-Style Modulation

Guanting Guo^{2†} Shenglong Hu^{2†} Kaihua Zhang^{1,3‡} Guangcan Liu³ Min Xia²

¹IEIT SYSTEMS Co., Ltd.

²School of Automation, Nanjing University of Information Science and Technology, Nanjing, China

³School of Automation, Southeast University, Nanjing, China

zhkhua@gmail.com

Abstract

The existing co-salient object detection (CoSOD) models lack sufficient generalization ability in real-world scenarios. To address this limitation, this paper presents CoMCS, a generalizable co-salient object detection framework that enhances robustness to unseen domains through mixed content-style modulation. The framework comprises a Mixed Content Modulator (MCM), a Mixed Style Modulator (MSM), and a Semantic Contrast Module (SCM), which collectively extract domain-invariant scene structure priors and augment source domain styles to bridge domain gaps. Specifically, we leverage CLIP to obtain multi-class semantic embeddings that capture conceptual knowledge of the entire scene. The MCM then models semantic relationships between co-salient object prototypes and these embeddings via cross-attention, capturing scene structures resilient to distribution shifts. Meanwhile, the MSM simulates domain perturbations by synthesizing feature statistics during training, enabling style augmentation within the source domain. To reduce representation ambiguity in unseen domains, the SCM employs a uniformity loss to ensure well-separated prototypes in hyperspherical space. Additionally, we construct an Unseen-domain Dataset (UND) with image groups from CoCA, CoSOD3k, and CoSal2015 to further evaluate generalization. Extensive experiments on four benchmarks demonstrate that CoMCS outperforms various state-of-the-art methods.

1. Introduction

Co-Salient Object Detection (CoSOD) aims at identifying common salient objects across a set of relevant images. It plays a crucial role in numerous real-world applications, such as object detection in surveillance footage [2, 22, 41], multi-camera object detection [20] and image search for

[†]Equal contribution. [‡]Corresponding author. This work is supported in part by the NSFC under Grant Nos. 62276141.

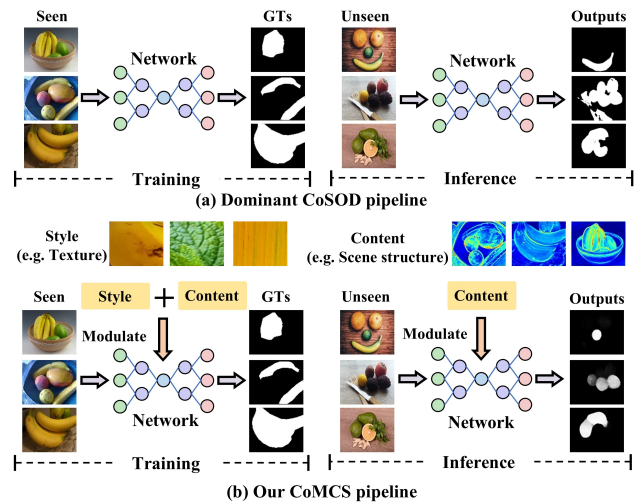


Figure 1. Comparison of CoSOD pipeline. (a) The dominant CoSOD framework [55, 64, 65] relies supervised learning on a pre-defined training dataset, struggling to generalize to unseen domains. (b) Our CoMCS improves generalization through joint content-style modulation to learn a robust representation, achieving more accurate predictions.

given images in search engines [26], to name a few.

Generally, existing leading CoSOD methods [14, 55, 60, 63–65] follow three steps to segment co-objects. First, a pre-trained encoder (e.g. VGG-16 [42] or PVT-V2 [50]) extracts multi-level image features, including deep features with rich content semantic information of target shapes or scene structures [5] and shallow features with rich style details of texture and color [16]. Subsequently, the network aggregates these features to derive co-object consensus with coupling content and style cues. Finally, the consensus cues are fed into a U-shaped [43, 60, 68] or FPN-structured [14, 55, 65] decoder, producing the co-salient object prediction. Among them, how to extract high-quality consensus cues is crucial to the model.

Despite the demonstrated success, existing works [43,

55, 63–65] exhibit certain limitations in consensus extraction due to the lack of generalization to unseen content and style. Firstly, the nature of supervised learning on a limited training set tends to produce a model that overfits to the semantic content present in the training data. When encountering real-world scenes with unseen categories, the trained model often suffers from poor generalization, leading to incorrect predictions. As illustrated by Figure 1 (a), due to the semantic overfitting issue, the model mistakenly treats the seen category (*bananas*) as the co-salient object of the unseen co-object (*avocado*). Secondly, due to the supervision from binary labels during training, the model tends to overfit some generic style features (e.g. texture) with similar patterns in the training dataset. When testing on unseen domains with a large domain gap to the source domain, the domain-specific style patterns in the consensus features will degrade the model to yield unsatisfying prediction quality.

Domain generalization (DG) [67] as an emerging research topic is an intuitive solution to the two problems. Several DG methods [9, 30, 66] increase the style diversity of training data by optimizing and augmenting the style features (e.g. texture, color and tone) of training data. Other DG methods [3, 19, 58] improve the generalization ability of the model in the face of domain perturbations in the data by simultaneously optimizing the content features (e.g. edge structure and object shape) and style features of the training data. Existing works [5, 21, 34, 36] have demonstrated that deep neural networks can encode not only content information but also style information. Although these methods achieve good results, due to the lack of perception of content information outside the training set, they cannot effectively improve the generalization ability of the model when faced with unseen categories and styles. At the same time, semantic content can be preserved while editing stylistic information. Since the content information and the style information can be encoded simultaneously, we enhance the generalization ability of the model by modulating the style information and the content information respectively.

In this paper, we propose a Mixed Content Style Modulation framework for generalizable CoSOD, dubbed CoMCS, which effectively addresses the distribution shift issue across domains. First, we utilize the CLIP model to derive conceptual knowledge linked to the semantic categories within the whole scene. This yields domain-invariant multi-class semantic embeddings, which are subsequently fed into the designed Mixed Content Modulator (MCM). Then, the MCM models the semantic interactions among the prototypes of co-salient objects and the multi-class semantic embeddings via a cross-attention mechanism [46]. This proficiently captures domain-invariant scene structure priors, which contribute to mitigating scene distribution shifts in previously unseen domains. Meanwhile, we design the Mixed Style Modulator (MSM) to address uncertainties

arising from domain shifts. During training, the MSM synthesizes feature statistics-specifically, the mean and standard deviation to generate novel stylistic variations, thereby enhancing data augmentation within the source domain. Finally, we design the collaborative Semantic Contrast Module (SCM) that enhances the co-features by inter-group attention. Meanwhile, the SCM utilizes a uniform loss function, ensuring that the learned prototypes are uniformly distributed within the hyperspherical space, thereby effectively mitigating the ambiguity associated with co-salient object representations from unseen domains.

Besides, to verify the superior generalization of the proposed CoMCS model to unseen domains. We construct an unseen-domain benchmark dataset (UND), which selects 50 co-object image groups with unseen categories during training from CoCA [63], CoSOD3k [13], CoSal2015 [57] with a total of 290 co-object categories. Extensive evaluations on CoCA [63], CoSOD3k [13], CoSal2015 [57] and the proposed UND demonstrate favorable performance of our CoMCS to a variety of state-of-the-art methods.

Our contributions are summarized as follows:

- A novel mixed content-style modulation framework, termed CoMCS, is introduced for generalizable CoSOD, which improves the model’s generalization ability across arbitrary unseen domains by systematically adjusting both style and content features
- An MCM is proposed to establish the association between the co-object prototype and the multi-class semantic embedding, which proficiently encapsulates the domain-invariant structural priors of scenes.
- An MSM is designed to increase the style diversity of training data by simulating various style characteristics.
- An SCM is built to capture the co-features and reduce the degree of ambiguity between different co-features.

2. Related Work

2.1. Co-Salient Object Detection

For the Co-Salient Object Detection (CoSOD) task, how to extract consensus information representing cooperative objects from a group of image features is very important for obtaining high quality prediction results. To extract consensus information, earlier work [40, 44, 52, 53, 57] employed hand-designed features or extracted consensus information from image groups through simple convolution and line projection operations. Due to inadequate exploration of interrelationships in feature maps, these methods can only extract low-quality consensus information. In order to improve the quality of the consensus information extraction, a lot of work [14, 15, 43, 60–62] has been done to extract higher quality consensus features through different network designs. [60, 61] use dynamic convolution to extract cooperative features from multi-scale feature maps. [43, 62] use

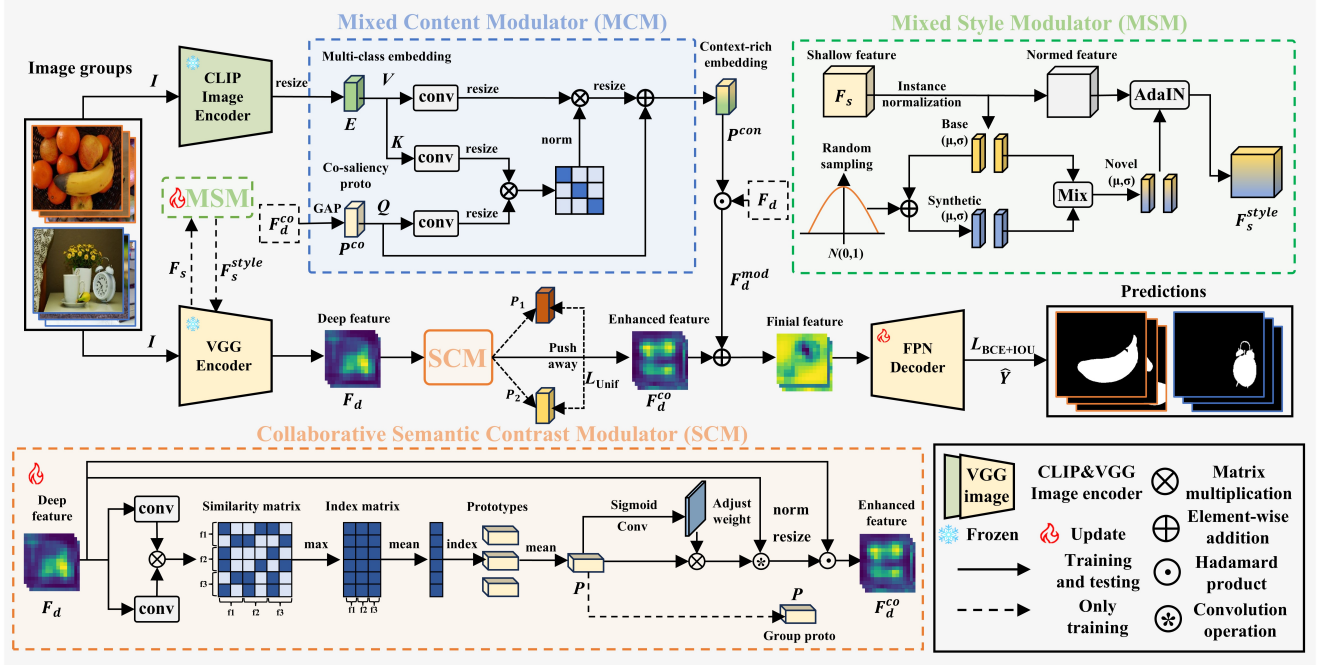


Figure 2. Pipeline of our CoMCS. The whole network consists of three main modules: MCM, MSM and SCM. Among them, MCM captures domain-invariant structural information by establishing collaborative object prototypes and multi-class semantic embeddings; MSM increases the style diversity of the training domain by simulating multiple styles; SCM is designed to capture cooperative features and reduce the interfering noise contained in them. Specifically, MSM is only used during training.

the self-attention mechanism to learn the cooperative expression from the feature map. [14, 15] use relevant attention mechanisms between image groups to find collaborative consensus. In addition, these methods [14, 55, 59, 63–65, 68] enhance the ability of the model to distinguish different cooperative objects and further obtain high-quality consensus information by simultaneously or separately utilizing special training modes and additional data sets, category labels and contrast learning. Among them, [59, 63] use additional classifiers to enhance the model’s semantic awareness of cooperative objects. [55, 65] use the idea of contrastive learning to push apart different categories of features as much as possible during training. With the aid of classifier, [14, 64] use inter-group training to enhance the ability of the network to distinguish cooperative features by processing two kinds of cooperative objects at the same time. [68] excludes non-cooperative features by iterative clustering under the guidance of additional data sets. Although these methods improve the quality of the model’s extraction of collaborative consensus. However, they ignore the impact of unseen class objects and out-of-domain data distribution in real application scenarios. Due to the lack of semantic awareness of unseen class objects and overfitting to domain-specific information, the consensus extracted by these methods sometimes contains non-cooperative objects incorrectly, resulting in incorrect prediction results. In this

paper, we propose the CoMCS framework to enhance the semantic perception of unseen class objects and improve the generalization ability of the model through the joint modulation of content and style.

2.2. Domain Generalization

Generalization of a model trained in the source domain to an environment with different data distributions is a challenging and long-studied task [48]. Among them, how to make the model learn the domain-invariant features is the main purpose of various generalization methods. These methods can be roughly divided into three categories. Feature extension [24, 48, 56, 58, 66], feature alignment [27, 29, 31, 35] and feature decoupling [5, 6, 36, 38]. Among them, feature alignment is to match the statistical features of the two domains in the training process or use contrast learning to narrow the semantic features extracted from the two domains as much as possible to enhance the model’s learning of invariant feature representations. Feature decoupling is to decompose the original feature into domain variation and domain invariant features and enhancing the generalization ability of the model to any domain by emphasizing the learning of domain invariant features. Feature extension is a simple and effective method, which extends the style of the training data at the image or feature level. By training the model on data with more diverse distribution, it can learn more in-

variant features independently, thus reducing the overfitting of the style features of the training domain.

3. Proposed Method

3.1. Method Overview

Figure 2 shows the overall architecture of our CoMCS. The CoMCS has two branches with a pre-trained CLIP image encoder [39] and a VGG encoder [42], respectively. During training, a set of collaborative images $\mathcal{I} = \{\mathcal{I}\}_{n=1}^N$ are fed into the model as one batch. First, the VGG encoder accepts $\mathcal{I}$ as input, outputting its corresponding multilevel features $\mathcal{F} = \{\mathcal{F}_s, \mathcal{F}_m, \mathcal{F}_d\}$, where $\mathcal{F}_s$ denotes the shallow features with rich style information, $\mathcal{F}_d$ denotes the deep features with rich content information and $\mathcal{F}_m$ denotes the other middle-stage features. Meanwhile, the CLIP image encoder inputs $\mathcal{I}$ to achieve a multi-class semantic embedding $\mathcal{E}$ that is domain-invariant to different scenes. Then, the shallow features $\mathcal{F}_s$ are sent to the MSM to synthesize the styles, getting the augmented shallow features $\mathcal{F}_s^{style}$. The deep features $\mathcal{F}_d$ are sent to the SCM, getting the co-saliency-enhanced features $\mathcal{F}_d^{co}$ followed by a global average pooling (GAP) layer [32], yielding the co-saliency prototype $\mathcal{P}^{co}$. Then, the embedding $\mathcal{E}$ and the prototype $\mathcal{P}^{co}$ are fed into the MCM, outputting the context-rich prototype $\mathcal{P}^{con}$. Next, $\mathcal{P}^{con}$ is used to modulate the deep features $\mathcal{F}_d$, yielding the modulated deep features $\mathcal{F}_d^{mod}$ with the rich scene structure information. Finally, we add $\mathcal{F}_d^{co}$ and $\mathcal{F}_d^{mod}$ element by element, and send them to the FPN decoder [33], yielding the final prediction. The total training process is supervised under a combination of a uniform loss L_{Uni} [49], a BCE loss L_{BCE} [7] and an IOU loss L_{IOU} [23]. During inference, a set of collaborative images will be fed into the model without MSM to obtain the final prediction.

3.2. Mixed Content Modulator

With the SCM, we get the co-saliency enhanced features $\mathcal{F}_d^{co} \in R^{N \times C \times H \times W}$, followed by the GAP [32] to obtain the co-saliency prototypes $\mathcal{P}^{co} \in R^{N \times C \times 1 \times 1}$. Meanwhile, through the CLIP image encoder, we obtain the multi-class embedding $\mathcal{E} \in R^{N \times C \times 1 \times 1}$, which contains domain-invariant semantic information across different domains. Next, the $\mathcal{P}^{co}$ and $\mathcal{E}$ are fed into a set of convolutional layers to generate the query $\mathcal{Q}$, key $\mathcal{K}$ and value $\mathcal{V}$. Among them, $\mathcal{Q}$ is from the co-saliency prototype $\mathcal{P}^{co}$, which captures seen co-saliency content information, while $\mathcal{K}$ and $\mathcal{V}$ are from the multi-class embedding $\mathcal{E}$, which encodes domain-invariant semantics from the whole scene. Subsequently, to learn the scene structure priors, the cross-attention mechanism [46] is employed to model the seman-

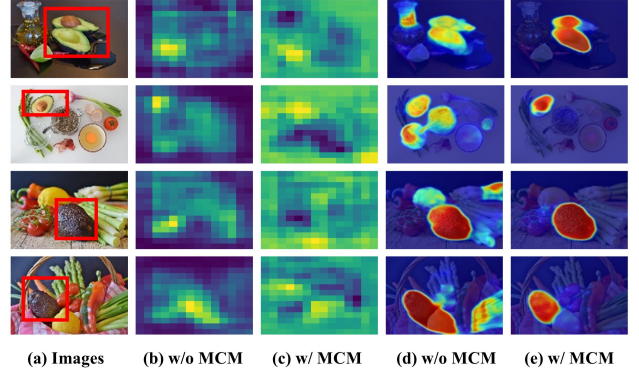


Figure 3. Visualization of the high-level feature maps (the 2^{nd} and 3^{rd} columns) and the activation maps (the 4^{th} and 5^{th} columns) of prediction layers with and without MCM. With MCM, the scene structures are highlighted that help to filter out more distractors, yielding more accurate predictions.

tic relationship between $\mathcal{E}$ and $\mathcal{P}^{co}$ as

$$\begin{cases} \text{Attention}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = \text{Softmax}\left(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{d_k}}\right)\mathcal{V}, \\ \mathcal{P}^{con} = \text{Attention}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) + \mathcal{Q}, \end{cases} \quad (1)$$

where $\text{Softmax}(\cdot)$ denotes the softmax function, $\top$ is the transpose operator, $\sqrt{d_k}$ is a scaling factor used to stabilize the gradients during the attention computation. The new context-rich prototype $\mathcal{P}^{con} \in R^{N \times C \times 1 \times 1}$ encodes the domain-invariant scene structure priors. Then, these new multi-class prototypes $\mathcal{P}^{con}$ are element-wise multiplied with the deep features $\mathcal{F}_d \in R^{N \times C \times H \times W}$, yielding the modulated features $\mathcal{F}^{mod} \in R^{N \times C \times H \times W}$ which effectively capture the scene structure priors.

Figure 3 shows the effect of the MCM. The semantic relationship between different classes should be domain-invariant, e.g. the unseen co-saliency object *avocado* and interference object *lemon* are closely related across domains. The model without MCM only considers part of content information, while the model with MCM considers more comprehensive scene structure information, thus improving the model's discrimination to co-objects.

3.3. Mixed Style Modulator

After passing through the first block of the VGG encoder, we get the shallow feature $\mathcal{F}_s \in R^{N \times C \times H \times W}$, where N denotes the number of feature maps, C represents the channel count of the features, and H , W denote the spatial height and width dimensions of the features, respectively. We then compute the mean and standard deviation across the spatial dimensions of each channel using

$$\begin{cases} \mu(\mathcal{F}_s) = \frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W \mathcal{F}_s, \\ \sigma(\mathcal{F}_s) = \sqrt{\frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W (\mathcal{F}_s - \mu(\mathcal{F}_s) + \epsilon)^2}, \end{cases} \quad (2)$$

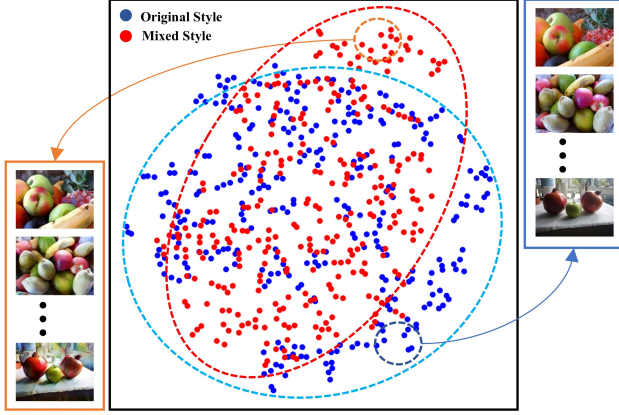


Figure 4. The t-SNE [45] visualization of the shallow feature statistics with the MSM, where novel domains are synthesized via generating feature statistics with diverse distribution possibilities. The images on the right are the original image, and the images on the left is the restored image using a pre-trained style transfer model [21].

where ϵ is a small constant added to prevent numerical instability due to zero variance respectively. These first-order (mean) and second-order (standard deviation) statistics collectively encapsulate the style information inherent in the shallow features. Since variance can represent the degree of dispersion of the data [30], we quantify the style dispersion across the two feature groups by computing their variances via

$$\begin{cases} S_\mu^2 = \frac{1}{N} \sum_{n=1}^N (\mu(\mathcal{F}_s) - (\frac{1}{N} \sum_{n=1}^N \mu(\mathcal{F}_s)))^2, \\ S_\sigma^2 = \frac{1}{N} \sum_{n=1}^N (\sigma(\mathcal{F}_s) - (\frac{1}{N} \sum_{n=1}^N \sigma(\mathcal{F}_s)))^2, \end{cases} \quad (3)$$

where $S_\mu^2 \in R^{1 \times C \times 1 \times 1}$ and $S_\sigma^2 \in R^{1 \times C \times 1 \times 1}$ denote the variance of the mean and standard deviation statistics across the batch, respectively. These metrics capture the dispersion of style parameters within the current batch. Next, we introduce randomized style statistics $\mu_{\text{random}}(\mathcal{F}_s)$ and $\sigma_{\text{random}}(\mathcal{F}_s)$ based on the obtained dispersion degree index as

$$\begin{cases} \mu_{\text{random}}(\mathcal{F}_s) = \mu(\mathcal{F}_s) + w\epsilon_\mu S_\mu^2, \epsilon_\mu \sim \mathcal{N}(0, 1), \\ \sigma_{\text{random}}(\mathcal{F}_s) = \sigma(\mathcal{F}_s) + w\epsilon_\sigma S_\sigma^2, \epsilon_\sigma \sim \mathcal{N}(0, 1), \end{cases} \quad (4)$$

where $w \in R^{N \times C \times 1 \times 1}$ is a scaling factor that controls the magnitude of the injected noise. Finally, we interpolate the randomized style $\{\mu_{\text{random}}(\mathcal{F}_s), \sigma_{\text{random}}(\mathcal{F}_s)\}$ with the original style $\{\mu(\mathcal{F}_s), \sigma(\mathcal{F}_s)\}$ using mixing coefficients λ , which are sampled from a Beta distribution $\lambda \in \text{Beta}(\alpha, \alpha)$. The combined style parameters are formalized as follows

$$\begin{cases} \gamma_{(new)} = \mu_{\text{random}}(\mathcal{F}_s) + (1 - \lambda)\mu(\mathcal{F}_s), \\ \beta_{(new)} = \sigma_{\text{random}}(\mathcal{F}_s) + (1 - \lambda)\sigma(\mathcal{F}_s). \end{cases} \quad (5)$$

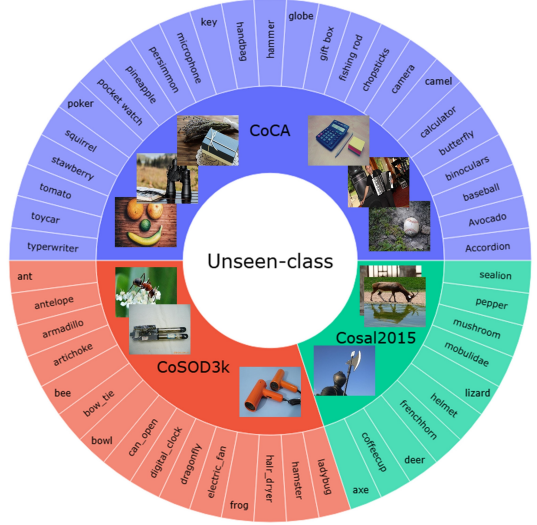


Figure 5. Category statistics of the proposed unseen-class dataset.

After getting the blended style, the AdaIN operation [21] is used to adjust the original feature maps as

$$MSM(x) = \beta_{(new)} \left(\frac{\mathcal{F}_s - \mu(\mathcal{F}_s)}{\sigma(\mathcal{F}_s)} \right) + \gamma_{(new)}. \quad (6)$$

Then, we obtain the styled features $\mathcal{F}_s^{style} \in R^{N \times C \times H \times W}$ with diverse styles. After that, the styled features $\mathcal{F}_s^{style}$ is fed into the VGG encoder. In the training process, we will choose whether to apply the new style with a random probability of p .

Figure 4 shows the effect of the MSM. It can be clearly seen that after the MSM modulates the original style features, the new style features extend the scope of the original style, thereby augmenting the training set domain that is helpful to improve the robustness of the trained model.

3.4. Collaborative Semantic Contrast Module

After passing the MSM module and the remaining blocks of the VGG encoder, we get the deep feature $\mathcal{F}_d \in R^{N \times C \times H \times W}$. Then, the convolutional layers are employed to generate two feature maps, which are subsequently reshaped to the size of $R^{NHW \times C}$. These reshaped feature maps are then utilized to compute a correlation map $S \in R^{NHW \times NHW}$, representing the similarity between images within the group. Subsequently, the correlation map S is reshaped to the size of $R^{NHW \times N \times HW}$. By identifying the vector with the highest score in the $H \times W$ dimension and averaging over the N dimension, we extract the channel vector that exhibits the highest similarity in the spatial dimension for each feature map. These vectors are regarded as primary prototypes, denoted as $\mathcal{P}^{pr} \in R^{N \times C \times 1 \times 1}$, which encapsulate collaborative information within the image groups. To further enhance the representation of coop-

erative information, the N dimension of the two prototypes is averaged. An adaptive adjustment operation is then applied on $\mathcal{P}^{pr}$ as follow

$$\mathcal{P}^{ad} = \mathcal{P}^{pr} \odot (\text{Sigmoid}(\text{Conv}(\mathcal{P}^{pr}))), \quad (7)$$

where $\odot$ denotes element-wise multiplication, Sigmoid represents the activation function, and Conv signifies a 1-dimensional convolutional operation. Subsequently, the adjusted collaborative prototypes $\mathcal{P}^{ad}$ are employed as convolutional kernels to perform convolution operations on the corresponding deep features $\mathcal{F}_d$ to obtain the collaborative features, which are then element-wise multiplied with the original deep feature maps $\mathcal{F}_d$ to produce enhanced collaborative feature maps $\mathcal{F}_d^{en}$.

3.5. Loss Function

Our loss function for training is a composite of three distinct loss functions: the uniform loss L_{Uni} [49], the BCE loss L_{BCE} [7], and the IOU loss L_{IOU} [23] as

$$L_{Uni}(P_1, P_2; t) = \log e^{-t(1 - \frac{P_1 \cdot P_2}{\|P_1\| \cdot \|P_2\|})}, \quad (8)$$

where t is the temperature coefficient, P_1 and P_2 are the semantic embeddings representing each group of collaborative features. $L_{Uni}(P_1, P_2; t)$ in (8) pushes P_1 and P_2 away from each other that ensures that the learned prototypes are uniformly distributed within the hyperspherical space [49].

The BCE loss L_{BCE} [7] and the IOU loss L_{IOU} [23] are defined as

$$\begin{cases} L_{BCE} = -\frac{1}{N} \sum_{i=1}^N [Y_i \log(\hat{Y}_i) + (1 - Y_i) \log(1 - \hat{Y}_i)], \\ L_{IOU} = 1 - \frac{1}{N} \sum_{i=1}^N \frac{|Y_i \cap \hat{Y}_i|}{|Y_i \cup \hat{Y}_i|}, \end{cases} \quad (9)$$

where Y is the ground-truth and $\hat{Y}$ is the predicted mask. Then the total loss function is

$$L_{all} = \lambda_1 L_{Uni} + \lambda_2 L_{BCE} + \lambda_3 L_{IOU}, \quad (10)$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyper-parameters that balance these three loss functions.

4. Experiments

4.1. Datasets and Evaluation Metrics

Training Sets For effectively training our CoMCS model, following other state-of-the-art CoSOD methods [43, 55, 64, 65], we combine the CoCo-SEG [47] and DUTS-class [63] as our training dataset.

Benchmark Test Sets We select three commonly-used CoSOD benchmark datasets to evaluate our proposed approach, including CoSal2015 [57], CoSOD3k [12] and CoCA [63].

Unseen-domain Test Sets To fully evaluate the generalization ability of our model in unseen domains, we construct a new dataset termed as unseen-domain benchmark dataset (UND). The UND selects 50 sets of cooperative images with unseen categories at the training sets (25 categories from CoCA, 15 categories from CoSOD3k and 10 categories from CoSal2015). Figure 5 shows the category statistics of our UND.

Evaluation Metrics We utilize four widely-used evaluation metrics, including S-measure S_α [10], Maximum E-measure E_ϕ^{max} [11], Maximum F-measure F_β^{max} [1], and MAE [4].

4.2. Implementation Details

Since most of mainstream methods such DCFM [55], GCoNet+ [64], and DMT [28] leverage VGG-16 backbone, for fair comparison, we also employ pre-trained VGG-16 [42] as an encoder for multiscale feature extraction and CLIP-ViT-B/32 [39] for semantic embedding extraction. The decoder architecture follows the FPN [33]. All images are resized to 224×224 pixels during both training and inference phases. During training, images are processed in two-category batches, while inference handles single-category image sets. The model is trained for 200 epochs using the Adam optimizer [25] with an initial learning rate of 1e-4, weight decay of 1e-4 and β parameters ($\beta_1=0.9, \beta_2=0.999$). All experiments are implemented in PyTorch [37] on a NVIDIA GeForce RTX 4090 GPU.

4.3. Comparison to State-of-The-Art Methods

Compared Methods To provide a comprehensive evaluation, we compare our CoMCS with 17 state-of-the-art methods over the past four years, including GICD [63], CoEGNet [13], DeepACG [59], CADC [60], GCoNet+ [64], DCFM [55], UFO [43], TCNet [15], CoRP [68], GCoNet [14], MCCL [65], CADC+ [61], ICSM [51], IPPO [54], ASCoD [17], E2HCoD [8], and UniTR [18].

Quantitative Results Table 1 lists the quantitative results on benchmark test sets, where our CoMCS model achieves superior performance in terms of all evaluation metrics. Specifically, on the most challenging CoCA dataset, our model achieves gains of 0.9% in S_α , 1.2% in E_ϕ^{max} and 0.9% in F_β^{max} compared to the second-best results.

Table 3 lists the quantitative results on unseen-domain test sets. Our CoMCS achieves the best performance in terms of all metrics, especially outperforming the second-best performing MCCL by 2.2% in S_α , the second-best performing GCoNet+ by 2.7% in F_β^{max} . These experiments validate the high generalization capability of our model to unseen domains.

Table 1. Quantitative comparison between our CoMCS and other state-of-the-art methods on three regular benchmark datasets. Red and blue fonts indicate best and second-best performance, respectively.

Methods	CoCA				CoSOD3k				CoSal2015			
	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\phi^{max} \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\phi^{max} \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\phi^{max} \uparrow$	$MAE \downarrow$
GICD(ECCV2020)	0.658	0.513	0.715	0.126	0.797	0.770	0.848	0.079	0.844	0.844	0.887	0.071
DeepACG(CVPR2021)	0.688	0.552	0.771	0.102	0.792	0.756	0.838	0.089	0.854	0.842	0.892	0.064
CADC(ICCV2021)	0.681	0.548	0.744	0.132	0.801	0.840	0.859	0.096	0.866	0.862	0.906	0.064
GCoNet(CVPR2021)	0.673	0.544	0.760	0.105	0.802	0.777	0.860	0.071	0.845	0.847	0.887	0.068
DCFM(CVPR2022)	0.710	0.598	0.783	0.085	0.810	0.805	0.874	0.067	0.838	0.856	0.892	0.067
CoEGNet(TPAMI2022)	0.612	0.493	0.717	0.106	0.762	0.736	0.825	0.092	0.836	0.832	0.882	0.077
UFO(TMM2023)	0.697	0.571	0.782	0.095	0.819	0.797	0.874	0.073	0.860	0.865	0.906	0.064
TCNet(TCSVT2023)	0.685	0.552	0.777	0.101	0.832	0.804	0.878	0.068	0.870	0.864	0.911	0.054
CADC+(PAMI2023)	0.708	0.609	0.791	0.107	0.823	0.808	0.876	0.070	0.875	0.889	0.922	0.060
CoRP(TPAMI2023)	0.703	0.575	0.741	0.110	0.820	0.794	0.864	0.075	0.859	0.864	0.896	0.060
MCCL(AAAI2023)	0.714	0.590	0.796	0.103	0.856	0.837	0.903	0.061	0.890	0.891	0.927	0.051
GCoNet+(PAMI2023)	0.738	0.637	0.812	0.081	0.843	0.834	0.901	0.062	0.881	0.891	0.924	0.056
UniTR(TMM2024)	0.708	0.574	0.789	0.089	0.853	0.834	0.900	0.072	0.860	0.857	0.906	0.063
ICSM(TiM2024)	0.725	0.626	0.806	0.099	0.853	0.843	0.890	0.061	0.890	0.893	0.924	0.052
IPPO(VRIH2024)	0.747	0.644	0.810	0.084	0.836	0.819	0.889	0.065	0.877	0.889	0.922	0.055
ASCoD(ICASSP2025)	0.738	0.640	0.816	0.083	0.830	0.819	0.900	0.063	0.860	0.867	0.907	0.061
E2HCoD(ICASSP2025)	0.729	0.626	0.805	0.081	0.837	0.820	0.886	0.071	0.880	0.877	0.912	0.055
CoMCS	0.747	0.649	0.821	0.084	0.857	0.843	0.907	0.058	0.890	0.896	0.929	0.052

Table 2. Quantitative ablation experiments on three modules of our CoMCS on three regular benchmark datasets.

Strategy	CoCA				CoSOD3k				CoSal2015			
	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\phi^{max} \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\phi^{max} \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\phi^{max} \uparrow$	$MAE \downarrow$
SCM	0.662	0.492	0.744	0.125	0.816	0.781	0.876	0.082	0.849	0.829	0.886	0.074
MCM	0.724	0.608	0.787	0.091	0.840	0.829	0.900	0.065	0.874	0.880	0.913	0.059
MSM	0.743	0.636	0.815	0.083	0.853	0.839	0.904	0.057	0.887	0.894	0.927	0.049
✓	0.735	0.629	0.800	0.093	0.851	0.840	0.907	0.063	0.884	0.889	0.921	0.056
✓	0.747	0.649	0.821	0.084	0.857	0.843	0.907	0.058	0.890	0.896	0.929	0.052

Table 3. The quantitative comparison between our CoMCS and other state-of-the-art methods is performed on unseen-domain benchmark datasets. Red and blue fonts indicate best and second-best performance, respectively.

Methods	Unseen-domain dataset			
	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\phi^{max} \uparrow$	$MAE \downarrow$
CADC(ICCV2021)	0.784	0.718	0.820	0.092
GCoNet(CVPR2021)	0.793	0.747	0.849	0.069
DCFM(CVPR2022)	0.783	0.743	0.846	0.071
UFO(TMM2023)	0.787	0.742	0.853	0.080
TCNet(TCSVT2023)	0.810	0.758	0.869	0.064
CoRP(TPAMI2023)	0.814	0.770	0.860	0.066
MCCL(AAAI2023)	0.834	0.789	0.886	0.060
GCoNet+(PAMI2023)	0.826	0.803	0.892	0.059
UniTR(TMM2024)	0.813	0.759	0.878	0.063
CoMCS	0.856	0.830	0.908	0.051

Table 4. The quantitative ablation experiments on three modules of our CoMCS model in unseen-domain dataset.

Strategy	Unseen-domain dataset			
	$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$E_\phi^{max} \uparrow$	$MAE \downarrow$
SCM	0.786	0.713	0.848	0.082
MCM	0.847	0.817	0.898	0.060
MSM	0.856	0.830	0.908	0.051

Qualitative Results Figure 6 show the qualitative comparison results of our CoMCS against other state-of-the-art methods. In the cooperative image groups containing nu-

merous interfering objects unrelated to the co-objects, our CoMCS effectively identifies the true co-objects while eliminating these interferences. For example, in the *avocado* group, the CoMCS precisely identifies and excludes non-cooperative objects such as *oranges* and *bananas*, which have highly similar appearances to the co-objects. In contrast, the other evaluated methods often produce erroneous predictions in the presence of such complex interferences.

4.4. Ablation Study

Effect of SCM As listed in Table 2, the SCM consistently improves the model performance in all three datasets. In CoCA, adding SCM achieves improvements of 6.2% in S_α , 11.6% in E_ϕ^{max} , 4.3% in F_β^{max} and 3.4% in MAE compared to the baseline model. In CoSOD3k, the respective gains are 2.4% in S_α , 4.8% in E_ϕ^{max} , 2.4% in F_β^{max} and 1.7% in MAE . In CoSal2015, the improvements are 2.5% in S_α , 5.1% in E_ϕ^{max} , 2.7% in F_β^{max} and 1.5% in MAE .

Figure 7 (a), (b) shows t-SNE visualization results of the baseline and the baseline with SCM, where adding SCM yields more compact semantic clusters for different classes.

Effect of MCM As listed in Table 2, the integration of the MCM significantly boosts model performance, particularly in CoCA, which has more interfering objects. Compared to using only the SCM, the addition of MCM yields im-

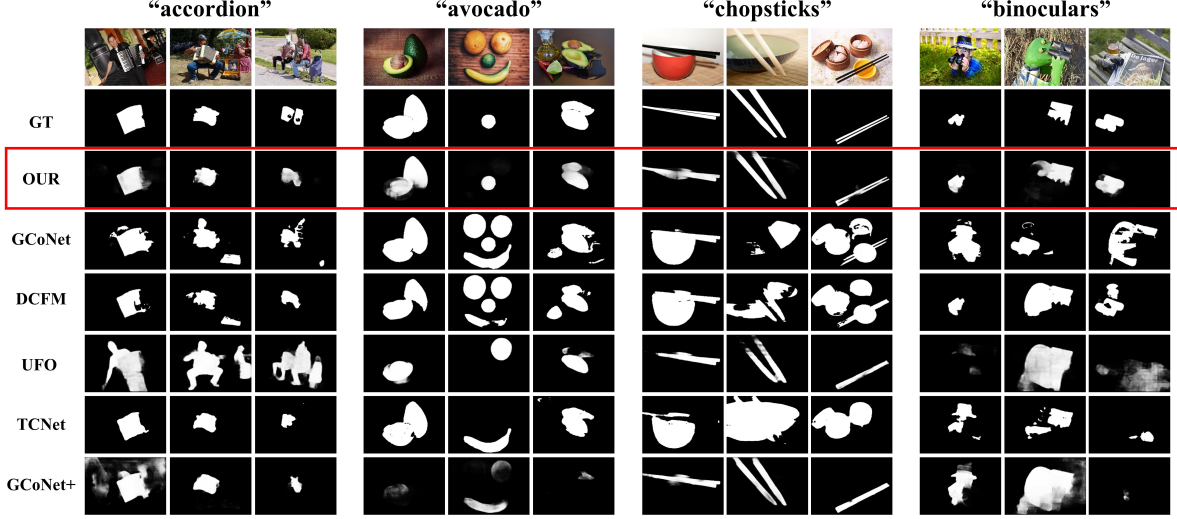


Figure 6. Qualitative comparison of our CoMCS with other methods on images containing objects of unseen class.

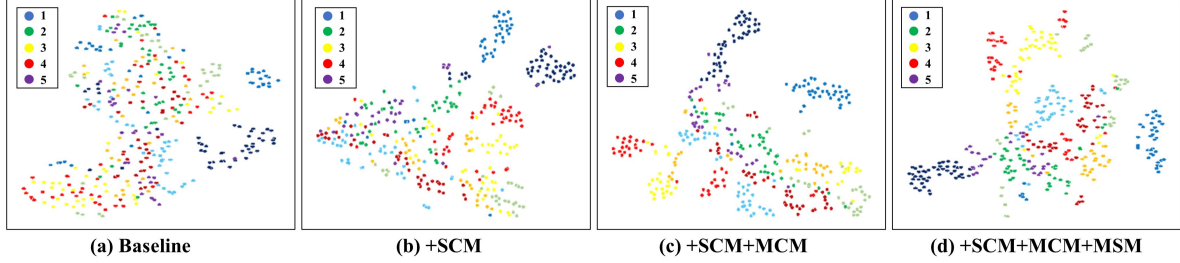


Figure 7. The t-SNE [45] visualization of high-level semantic features.

provements of 1.9% in S_α , 2.8% in F_β^{max} , 2.8% in E_ϕ^{max} and 3.4% in MAE . In the relatively simpler CoSOD3k, the MCM achieves respective gains of 1.3% in S_α , 1.0% in E_ϕ^{max} , 0.4% in F_β^{max} , and 1.7% in MAE . In the CoSal2015 dataset, which has fewer distractors, the SCM module enhances the model’s focus on co-objects, resulting in improvements of 1.3% in S_α , 1.4% in E_ϕ^{max} , 1.4% in F_β^{max} , and 1.0% in MAE .

As shown by Figure 7 (c), adding MCM further yields more semantic clusters than that without MCM as shown by Figure 7 (b), which verifies the effectiveness of the MCM.

Effect of MSM As listed in Table 2, the MSM consistently enhances model performance across all benchmark datasets. On the challenging CoCA, adding MSM achieves notable improvements of 1.1% in S_α , 2.1% in F_β^{max} , and 1.3% in E_ϕ^{max} compared to using only SCM and MCM, despite a marginal 0.2% MAE reduction. Similar performance gains are observed on CoSOD3k, with improvements of 1.1% in S_α , 1.1% in E_ϕ^{max} , 0.9% in F_β^{max} , and a slight 0.2% in MAE . Notably, even on the less com-

plex CoSal2015 with fewer distractors, the MSM maintains consistent improvements of 1.0% S_α , 0.9% E_ϕ^{max} , 0.8% in F_β^{max} , and 0.3% in MAE .

As shown by Figure 7 (d), all the semantic clusters are compact, thereby further validating the effectiveness of the CoMCS that excels at discriminating the co-salient objects from background distractors.

5. Conclusion

In this paper, we propose CoMCS, a novel framework for generalizable co-salient object detection that effectively mitigates cross-domain distribution shifts through mixed content and style modulation. The framework captures domain-invariant scene structures by modeling semantic interactions between co-salient object prototypes and multi-class embeddings, synthesizes diverse stylistic variations via feature statistics to enhance intra-domain augmentation, and disambiguates co-features through a dedicated co-feature modeling module. We further introduced a new dataset UND for better evaluate generalization. Extensive experiments demonstrate that CoMCS consistently outperforms state-of-the-art methods across all benchmarks.

References

- [1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1597–1604. IEEE, 2009. 6
- [2] J Arunnehru et al. Deep learning-based real-world object detection and improved anomaly detection for surveillance videos. *Materials Today: Proceedings*, 80:2911–2916, 2023. 1
- [3] Chaoqi Chen, Luyao Tang, Feng Liu, Gangming Zhao, Yue Huang, and Yizhou Yu. Mix and reason: Reasoning over semantic topology with data mixing for domain generalization. *Advances in Neural Information Processing Systems*, 35:33302–33315, 2022. 2
- [4] Ming-Ming Cheng, Jonathan Warrell, Wen-Yan Lin, Shuai Zheng, Vibhav Vineet, and Nigel Crook. Efficient salient region detection with soft image abstraction. In *Proceedings of the IEEE International Conference on Computer vision*, pages 1529–1536, 2013. 6
- [5] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11580–11590, 2021. 1, 2, 3
- [6] Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6614–6630, 2021. 3
- [7] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134:19–67, 2005. 4, 6
- [8] Chuang Ding, Zhidong Han, Gang Dong, Lingyan Liang, Dongchao Wen, and Kaihua Zhang. Easy-to-hard instance-level feature fusion for co-saliency detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 6
- [9] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul De Charette. A simple recipe for language-guided domain generalized segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23428–23437, 2024. 2
- [10] 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 international conference on computer vision*, pages 4548–4557, 2017. 6
- [11] 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
- [12] 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. 6
- [13] 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*, 44(8): 4339–4354, 2021. 2, 6
- [14] 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, 3, 6
- [15] Yanliang Ge, Qiao Zhang, Tian-Zhu Xiang, Cong Zhang, and Hongbo Bi. Tcnet: Co-salient object detection via parallel interaction of transformers and cnns. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(6):2600–2615, 2022. 2, 3, 6
- [16] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018. 1
- [17] Guanting Guo, Shenglong Hu, Huihui Song, and Kaihua Zhang. Learning joint appearance and shape co-representations for co-saliency detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 6
- [18] Ruohao Guo, Xianghua Ying, Yanyu Qi, and Liao Qu. Unitr: A unified transformer-based framework for co-object and multi-modal saliency detection. *IEEE transactions on multimedia*, 2024. 6
- [19] Minui Hong, Jinwoo Choi, and Gunhee Kim. Stylemix: Separating content and style for enhanced data augmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14862–14870, 2021. 2
- [20] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 1
- [21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2, 5
- [22] Zihao Jia, Shengkun Sun, Guangcan Liu, and Bo Liu. Mssd: multi-scale self-distillation for object detection. *Visual Intelligence*, 2(1):8, 2024. 1
- [23] 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. 4, 6
- [24] Alexander Jung. Imgaug documentation. *Readthedocs.io*, Jun, 25, 2019. 3
- [25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [26] Afshan Latif, Aqsa Rasheed, Umer Sajid, Jameel Ahmed, Nouman Ali, Naeem Iqbal Ratyal, Bushra Zafar, Saadat Hanif Dar, Muhammad Sajid, and Tehmina Khalil.

- Content-based image retrieval and feature extraction: A comprehensive review. *Mathematical problems in engineering*, 2019(1):9658350, 2019. 1
- [27] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9936–9946, 2022. 3
- [28] Long Li, Junwei Han, Ni Zhang, Nian Liu, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, and Fahad Shahbaz Khan. Discriminative co-saliency and background mining transformer for co-salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7247–7256, 2023. 6
- [29] Shuang Li, Shiji Song, Gao Huang, Zhengming Ding, and Cheng Wu. Domain invariant and class discriminative feature learning for visual domain adaptation. *IEEE transactions on image processing*, 27(9):4260–4273, 2018. 3
- [30] Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, and Ling-Yu Duan. Uncertainty modeling for out-of-distribution generalization. *arXiv preprint arXiv:2202.03958*, 2022. 2, 5
- [31] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018. 3
- [32] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 4
- [33] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4, 6
- [34] Chang Liu, Xudong Jiang, and Henghui Ding. Primitivenet: decomposing the global constraints for referring segmentation. *Visual Intelligence*, 2(1):16, 2024. 2
- [35] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11749–11756, 2020. 3
- [36] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the european conference on computer vision (ECCV)*, pages 464–479, 2018. 2, 3
- [37] A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. 6
- [38] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *International conference on machine learning*, pages 7728–7738. PMLR, 2020. 3
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 4, 6
- [40] Jingru Ren, Zhi Liu, Xiaofei Zhou, Cong Bai, and Guangling Sun. Co-saliency detection via integration of multi-layer convolutional features and inter-image propagation. *Neuro-computing*, 371:137–146, 2020. 2
- [41] Tomás Santos, Hélder Oliveira, and António Cunha. Systematic review on weapon detection in surveillance footage through deep learning. *Computer science review*, 51:100612, 2024. 1
- [42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 4, 6
- [43] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, Hanjing Su, and Qingyao Wu. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. *IEEE Transactions on Multimedia*, 26:313–325, 2023. 1, 2, 6
- [44] Chung-Chi Tsai, Weizhi Li, Kuang-Jui Hsu, Xiaoning Qian, and Yen-Yu Lin. Image co-saliency detection and co-segmentation via progressive joint optimization. *IEEE Transactions on Image Processing*, 28(1):56–71, 2018. 2
- [45] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 5, 8
- [46] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2, 4
- [47] 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*, pages 8917–8924, 2019. 6
- [48] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022. 3
- [49] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020. 4, 6
- [50] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational visual media*, 8(3):415–424, 2022. 1
- [51] Longsheng Wei and Jiu Huang. An information calibration and sliding mining network for co-saliency object detection. *IEEE Transactions on Instrumentation and Measurement*, 2024. 6
- [52] 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. 2
- [53] Lina Wei, Shanshan Zhao, Omar El Farouk Bourahla, Xi Li, Fei Wu, and Yueting Zhuang. Deep group-wise fully convolutional network for co-saliency detection with graph propagation. *IEEE Transactions on Image Processing*, 28(10):5052–5063, 2019. 2
- [54] Yang Wen, Yuhuan Wang, Hao Wang, Wuzhen Shi, and Wenming Cao. Co-salient object detection with iterative purification and predictive optimization. *Virtual Reality & Intelligent Hardware*, 6(5):396–407, 2024. 6

- [55] 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](#), [2](#), [3](#), [6](#)
- [56] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. [3](#)
- [57] Dingwen Zhang, Junwei Han, Chao Li, and Jingdong Wang. Co-saliency detection via looking deep and wide. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2994–3002, 2015. [2](#), [6](#)
- [58] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. [2](#), [3](#)
- [59] 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. [3](#), [6](#)
- [60] 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. [1](#), [2](#), [6](#)
- [61] Ni Zhang, Nian Liu, Fang Nan, and Junwei Han. Cad++: Advanced consensus-aware dynamic convolution for co-salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. [2](#), [6](#)
- [62] Qijian Zhang, Runmin Cong, Junhui Hou, Chongyi Li, and Yao Zhao. Coadnet: Collaborative aggregation-and-distribution networks for co-salient object detection. *Advances in neural information processing systems*, 33:6959–6970, 2020. [2](#)
- [63] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 455–472. Springer, 2020. [1](#), [2](#), [3](#), [6](#)
- [64] Peng Zheng, Huazhu Fu, Deng-Ping Fan, Qi Fan, Jie Qin, Yu-Wing Tai, Chi-Keung Tang, and Luc Van Gool. Gconet+: A stronger group collaborative co-salient object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10929–10946, 2023. [1](#), [3](#), [6](#)
- [65] Peng Zheng, Jie Qin, Shuo Wang, Tian-Zhu Xiang, and Huan Xiong. Memory-aided contrastive consensus learning for co-salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3687–3695, 2023. [1](#), [2](#), [3](#), [6](#)
- [66] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. [2](#), [3](#)
- [67] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4396–4415, 2022. [2](#)
- [68] Ziyue Zhu, Zhao Zhang, Zheng Lin, Xing Sun, and Ming-Ming Cheng. Co-salient object detection with co-representation purification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8193–8205, 2023. [1](#), [3](#), [6](#)