# Self-Supervised Co-Salient Object Detection via Unified Multi-Granularity Feature Learning

Mao Yuan[1,2], Guohua Lv[1,2,3*(✉)], Guangxiao Ma[4], and Zhengyang Zhang[1,2]

[1] Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
{10431230012,10431240210}@stu.qlu.edu.cn, guohualv@qlu.edu.cn
[2] Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan, China
[3] Shandong Key Laboratory of Ubiquitous Intelligent Computing, Jinan, China
[4] College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, China; mgx@sdust.edu.cn

**Abstract.** Co-salient object detection (CoSOD) seeks to identify common salient objects across related image groups, traditionally relying on supervised learning with extensive manual annotations. To address this challenge, we propose UMS-CoSOD, a novel self-supervised framework that eliminates the need for labeled data while achieving state-of-the-art performance. Our approach unfolds in three stages: cross-modal pseudo-label generation, feature learning and saliency prediction, and result refinement and boundary enhancement(RRBE). Initially, high-quality pseudo-labels are generated via a cross-modal self-supervised strategy, integrating instance segmentation masks and self-attention maps from different modalities of self-supervised information. The core feature learning stage employs an adaptive feature mixer module(AFMM) for intelligent fusion of initial features, a Comprehensive Feature Mining Module (CFMM) to capture inter-image co-salient cues, a Feature Refinement and Amplification Module (FRAM) for feature enhancement. During inference, Adaptive Threshold Binarization, Region-Level Feature Refinement and Dense CRFs are used to obtain correct and clearly bounded results. Evaluated on benchmark datasets (CoCA, CoSOD3k and CoSal2015), UMS-CoSOD significantly outperforms unsupervised methods, reducing MAE by 19.8% on CoSOD3k. It also achieves comparable performance to top-tier fully supervised methods. This framework offers a practical, annotation-free solution for CoSOD, with broad applicability in real-world scenarios.The code is available at UMS-CoSOD.

**Keywords:** Co-Salient Object Detection · Self-supervised Learning · Deep Learning · Vision Transformer.

## 1 Introduction

Co-salient object detection (CoSOD) aims to identify and segment objects that are consistently salient across a group of related images [6, 18, 24], leveraging

* Corresponding author: Guohua Lv (guohualv@qlu.edu.cn).

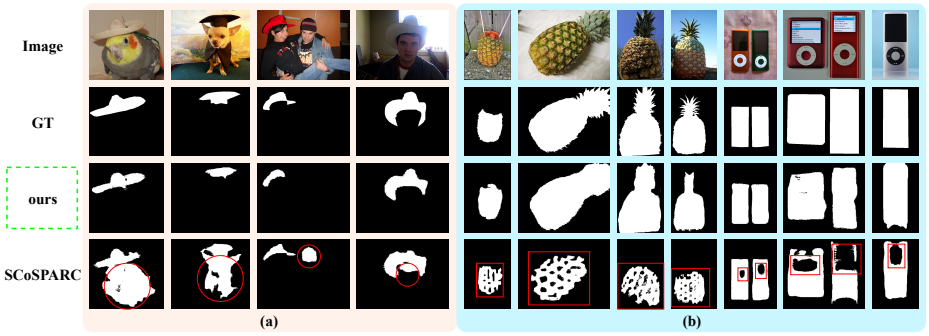


**Fig. 1.** Our method is compared with SCoSPARC. (a) and (b) respectively show instances of SCoSPARC's misidentification and Hole erosion phenomenon.

inter-image relationships to distinguish it from single-image salient object detection (SOD) [9, 14, 29, 35]. This capability is invaluable for applications such as image editing, video summarization, and robotic perception. While supervised CoSOD methods, such as GCoNet+ [34] and Visual Consensus Prompting [24], achieve high accuracy, their reliance on extensive manual annotations limits scalability. Unsupervised approaches, particularly self-supervised CoSOD, offer a compelling alternative by learning co-saliency from the inherent structure of image groups without explicit supervision, enabling broader real-world deployment [5, 6]. Recent CoSOD methods often adopt consensus extraction and dispersion paradigms, using techniques like Transformer-based inter-image associations [28] or similarity-based feature optimization [22].

Concurrently, Self-supervised learning (SSL) has revolutionized computer vision by enabling models to learn powerful visual representations without requiring manual annotations. One prominent approach in this domain is DINO [4] (self-Distillation with No Labels), which leverages Vision Transformers (ViTs) to extract semantically rich features from images. DINO distinguishes itself by learning discriminative patch-level representations through a teacher-student framework, where the student mimics the teacher's outputs across augmented views of the same image. This method has shown to naturally encode object-level semantics and performs remarkably well in tasks like image retrieval and segmentation without fine-tuning [4]. Comparative studies also find that DINO's contrastive learning produces more generalizable and transferable features than masked autoencoders [21].

However, applying SSL to CoSOD reveals challenges: existing self-supervised methods, such as those relying on patch-level correspondences [6], struggle with multiple salient objects or visually prominent distractors, often misinterpreting them as co-salient (see Fig. 1(a)). Additionally, phenomena like "hole erosion" [6] in predictions (Fig. 1(b))—where fine-grained details are lost—highlight limitations in capturing nuanced group-level commonalities amidst complex backgrounds.

To address these issues, we propose the Self-Supervised Co-Salient Object Detection via Unified Multi-Granularity Feature Learning (UMS-CoSOD) framework. Our approach eliminates the need for manual annotations by leveraging cross-modal self-supervised pseudo-label generation, integrating instance segmentation masks (e.g., HASSOD [3]) and self-attention maps from pre-trained Vision Transformers (e.g., DINO [4]). UMS-CoSOD operates in three stages:

1) Pseudo-Label Generation: A Cross-Modal Self-Supervised Pseudo-Label Generation (CM-SPLG) method synthesizes high-quality training data.

2) Feature Learning and Prediction: An Adaptive Feature Mixer (AFM) fuses multi-granularity features from DINO variants, followed by a Comprehensive Feature Mining Module (CFMM) for group-level consensus and a Feature Refinement and Amplification Module (FRAM) to enhance discriminability.

3) Refinement: Adaptive thresholding and Dense Conditional Random Fields sharpen boundaries and eliminate irrelevant regions. Our key contributions include:

- Cross-Modal Self-Supervised Pseudo-Label Generation: A novel mechanism to synthesize high-fidelity training sets by fusing diverse self-supervised visual cues.
- Feature Learning and Saliency Prediction: Our end-to-end trainable network is designed with dedicated modules (AFM, CFMM, FRAM) that progressively learn and refine co-saliency features from multiple scales and semantic levels, enabling robust handling of complex scenarios and object diversity. During inference, Adaptive Threshold Binarization, Region-Level Feature Refinement and Dense CRFs are used to obtain correct and clearly bounded results.
- State-of-the-Art Performance: UMS-CoSOD surpasses existing self-supervised methods and rivals top supervised approaches on benchmarks, without manual annotations.

## 2 Method

This section details our Self-Supervised Co-Salient Object Detection via Unified Multi-Granularity Feature Learning (UMS-CoSOD) framework, operating in three principal stages (Fig. 2): Stage 1: Pseudo-Label Generation. High-quality pseudo-ground truth maps are created via a cross-modal self-supervised method. Stage 2: Feature Learning and Saliency Prediction. This core training phase uses generated pseudo-labels and includes the AFMM, CFMM, and FRAM to obtain preliminary prediction results. Stage 3: Result Refinement and Boundary Enhancement (RRBE). An optional inference-only stage, it filters non-salient regions and optimizes boundaries.

### 2.1 Cross-modal Self-supervised Pseudo-Label Generation

In Stage 1, we introduce Cross-Modal Self-Supervised Pseudo-Label Generation (CM-SPLG) to create high-quality pseudo-label for CoSOD without man-

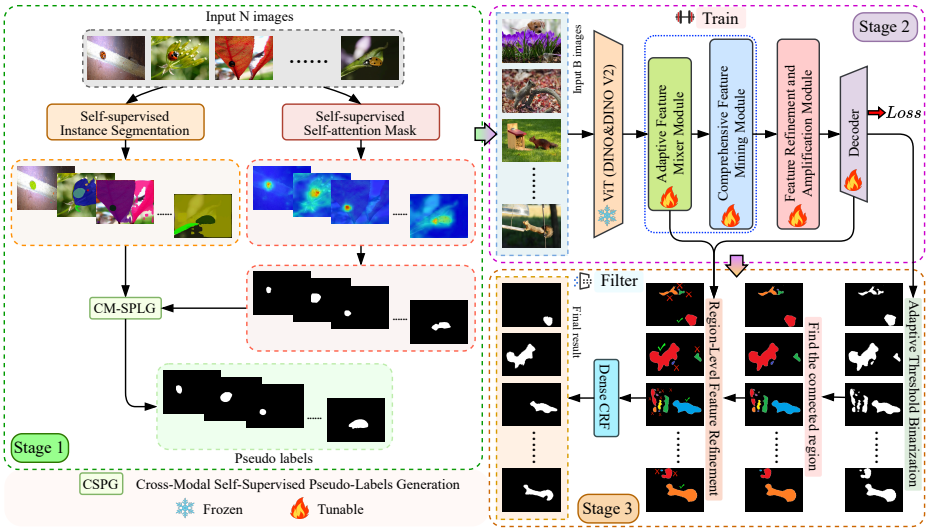


**Fig. 2.** Our proposed method operates in three stages: the first generates pseudo ground truth, the second is a training phase to produce preliminary results, and the third refines these results by eliminating non-salient and enhancing object boundaries.

ual annotations. CM-SPLG leverages co-occurrence information from two self-supervised models:

1) DINO: We extract self-attention maps from DINO's last transformer layer. These maps highlight salient foreground regions.

2) HASSOD: We obtain instance segmentation masks from HASSOD, which adaptively segments multi-scale objects. These masks provide more detailed object candidate regions than DINO.

To generate the final pseudo-label, we fuse information from both modalities using a tailored filtering algorithm:

1) Calculate the co-occurrence frequency $f^i$ of each category in the image group $G$ and sort them in descending order.

2) Sort the categories in each image in descending order according to the order of $f$ and take the first k categories.

3) From these candidates, we choose the pseudo-co-saliency category ($c^i_{cs_j}$) based on the overlap score $O^i_j$ between its HASSOD instance mask and the DINO self-attention map. The overlap score is calculated as:

$$O^i_j = \frac{Area(HM^i_j \cap SA^i)}{Area(HM^i_j \cup SA^i)}, \tag{1}$$

where $\cap$ represents the intersection, $\cup$ represents the union, and $Area$ refers to the area of the intersection or union, $i, j = \{1, 2, \cdots, n\}$, $HM^i_j$ is the mask of the $j$th category in each image, and $SA^i$ is the self-attention map of DINO for each image.

4) Finally, the HASSOD mask $HM_j^i$ corresponding to $c_{cs_j}^i$ becomes the pseudo-ground truth mask for that image.

This process ensures that the generated pseudo-masks accurately represent frequently co-occurring salient objects. The final pseudo-label Y are thus obtained.

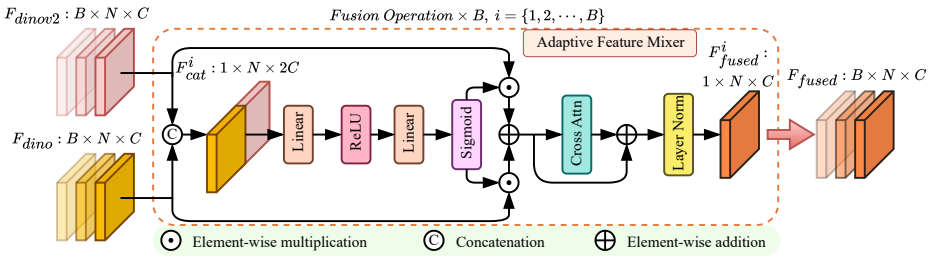


**Fig. 3.** The detailed structure of the Adaptive Feature Mixer module.

## 2.2 Feature Learning and Saliency Prediction

**Adaptive Feature Mixer Module** Our approach enhances CoSOD by combining features from pre-trained DINO and DINOv2 models, leveraging their distinct semantic capture capabilities for a richer feature space. This early-stage feature aggregation serves as a robust input for CoSOD processing. Our Adaptive Feature Mixer Module (AFMM) (Fig. 3) aggregates complementary semantic cues from DINO and DINOv2 backbones, providing a richer initial feature representation. Experiments show DINO excels in general feature extraction, while DINOv2 performs better in complex scenes with significant background noise. AFMM adaptively fuses their strengths, improving feature resilience and discriminative power.

For each input image $I_i$, we extract patch token features from both DINO ($F_{dino}^i \in \mathbb{R}^{1\times N\times C}$) and DINOv2 ($F_{dinov2}^i \in \mathbb{R}^{1\times N\times C}$). $F_{dinov2}^i$ is upsampled to match $F_{dino}^i$'s size. We then perform dynamic weighted fusion:

1) Concatenate $F_{dino}^i$ and $F_{dinov2}^i$ to form $F_{cat}^i \in \mathbb{R}^{1\times N\times 2C}$.

2) Feed $F_{cat}^i$ into a lightweight weight generation network (MLP with Linear, ReLU, Linear, Sigmoid) to output a per-patch dynamic weight $w_j \in [0, 1]$.

3) Compute the fused feature $F_{dyn}^i$ as:

$$F_{dyn}^i = w_j \odot F_{dino}^i + (1 - w_j) \odot F_{dinov2}^i, \tag{2}$$

where $\odot$ denotes element-wise multiplication.

This adaptive weighting intelligently leverages the stronger signal from either DINO or DINOv2 for different image regions.

The dynamically fused feature $F_{dyn}^i$ is further refined by a cross-modal attention mechanism (Multi-head Attention with C dimensions, 4 heads). $F_{dyn}^i$ serves as *query*, *key*, and *value*. The output, $A_{attn}^i$, represents self-attention enhanced features. Then after a residual connection and layer normalization, we

get $F^i_{fused} \in \mathbb{R}^{1\times N\times C}$, and these normalized features are concatenated along the batch dimension to obtain $F_{fused} \in \mathbb{R}^{B\times N\times C}$. This provides a rich, contextually aware feature set for effective co-saliency detection.

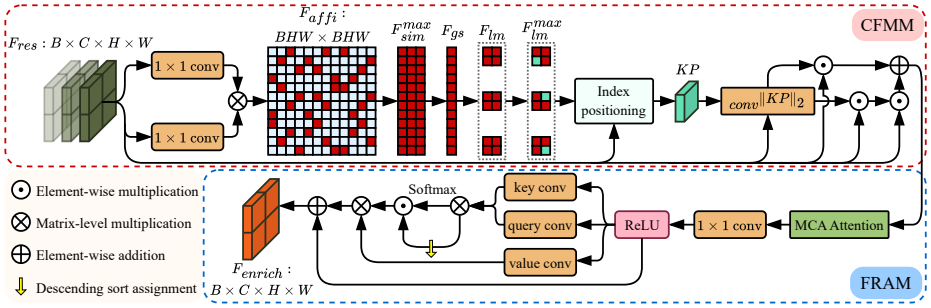


**Fig. 4.** Detailed structure of Comprehensive Feature Mining Module and Feature Refinement and Amplification Module.

**Comprehensive Feature Mining Module** As shown in Fig. 4, the Comprehensive Feature Mining Module (CFMM) extracts and refines group-level co-salient features for robust representation.

Initially, $F_{fused}$ is reshaped to $\mathbb{R}^{B\times C\times H\times W}$ to generate residual features $F_{res} \in \mathbb{R}^{B\times C\times H\times W}$. To uncover pixel-level collaborative patterns, a self-attention inspired mechanism is employed. $F_{res}$ is transformed via two $1\times 1$ convolutions into query ($F_{query}$) and key ($F_{key}$) features. These are reshaped and multiplied to generate a group-wise feature affinity map $F_{affi} \in \mathbb{R}^{BHW\times BHW}$, quantifying pixel similarity across the group. $F_{affi}$ is then reshaped to $\mathbb{R}^{BHW\times B\times HW}$.

The module then pinpoints co-salient locations. The maximum similarity value across all images for each pixel yields $F^{max}_{sim} \in \mathbb{R}^{BHW\times B}$. Averaging along rows of $F^{max}_{sim}$ gives a global group-level pixel importance score $F_{gs} \in \mathbb{R}^{BHW}$, spatially reshaped to $F_{lm} \in \mathbb{R}^{B\times H\times W}$. A group-level prototype vector, KeyPoints ($KP \in \mathbb{R}^{B\times C}$), is derived by retrieving the feature vector from $F_{res}$ corresponding to the maximum value in $F_{lm}$.

To enrich $KP$, a response map is computed. $F_{res}$ and $KP$ are Euclidean normalized ($\|F_{res}\|_2$, $\|KP\|_2$). $\|KP\|_2$ is convolved with $\|F_{res}\|_2$ to produce a multi-layer response map $RM \in \mathbb{R}^{B\times B\times H\times W}$. These maps are averaged to form a mean response map for each image, $RM_{mean} \in \mathbb{R}^{B\times H\times W}$, ensuring all pixels contribute. Next, element-wise multiplication between $RM_{mean}$ and $F_{res}$ is averaged to obtain the comprehensive feature $F_{comp} \in \mathbb{R}^{1\times C}$. Finally, we perform a fusion of $RM_{mean}$, $F_{comp}$, and $F_{res}$ to create $F_{merged} \in \mathbb{R}^{B\times C\times H\times W}$:

$$F_{merged} = F_{comp} \odot F_{res} + RM_{mean} \odot F_{res}, \tag{3}$$

where $F_{comp}$ and $RM_{mean}$ are broadcast to match $F_{res}$'s spatial dimensions before element-wise multiplication, integrating global context, local responses, and original features.

**Feature Refinement and Amplification Module** As illustrated in Fig. 4, the Feature Refinement and Amplification Module(FRAM) enhances co-salient information from the CFMM to improve object delineation. First, we incorporate an adaptive cross-scale attention mechanism (MCA [30]) into FRAM to capture multi-scale and multi-modal correlations. $F_{merged}$ is fed into this attention module, followed by a $1 \times 1$ convolution and ReLU activation, yielding $F_{act} \in \mathbb{R}^{N \times C \times H \times W}$.

To construct an enhanced attention map, $F_{act}$ is projected into Query ($F_{query}^{proj}$), Key ($F_{key}^{proj}$), and Value ($F_{value}^{proj}$) features. The raw attention map $A_{raw} \in \mathbb{R}^{B \times HW \times HW}$ is generated by multiplying $F_{key}^{proj}$ and $F_{query}^{proj}$. Positive attention signals in $A_{raw}$ are non-linearly amplified using an amplification matrix $S_{amp}$ (derived by sorting $A_{raw}$ and assigning larger indices to smaller values):

$$A_{amp}^{i,j} = \begin{cases} (S_{amp}^{i,j} + 1)^{\alpha}, & \text{if } A_{raw}^{i,j} > 0 \\ 1, & \text{else } A_{raw}^{i,j} \leq 0 \end{cases}, \tag{4}$$

where $\alpha$ is a tunable amplification factor.

$A_{raw}$ is also passed through a softmax function to obtain $A_{norm} \in \mathbb{R}^{B \times HW \times HW}$. The final enhanced attention map $A_{final} \in \mathbb{R}^{B \times HW \times HW}$ is derived by element-wise multiplication of $A_{norm}$ and $A_{amp}$.

Finally, the ultimate enhanced feature $F_{enrich} \in \mathbb{R}^{C \times H \times W}$ is obtained by combining $F_{value}^{proj}$ with $A_{final}$ and a residual connection with $F_{act}$:

$$F_{enrich} = F_{value}^{proj} \otimes A_{final} + F_{act}, \tag{5}$$

where $\otimes$ denotes matrix multiplication.

This refinement amplifies salient features, providing a highly discriminative representation for precise co-salient object detection. $F_{enrich}$ is then passed to the decoder for initial prediction map $M$ generation.

### 2.3 Result Refinement and Boundary Enhancement

**Adaptive Threshold Binarization** To generate precise binary segmentation masks from the predicted co-saliency maps ($M_i$), we apply an adaptive thresholding strategy. This dynamically adjusts the binarization threshold based on the confidence of the detected co-salient regions. A higher confidence map uses a lower threshold, while a less confident map requires a higher threshold to filter noise. The average per-pixel confidence, denoted as $conf$, for each map $M_i$ is computed by taking the sum of intensity values ($M_p$) for all confident pixels ($p$) where the pixel intensity is greater than or equal to the overall average intensity of the map ($\overline{M}$), and then dividing this sum by the total count of confident pixels ($N_c$). The adaptive threshold ($th$) is then determined by:

$$th = th_0 + \alpha_c(B_M - \overline{B_M}), \tag{6}$$

where $th_0$ is a baseline offset, $\alpha_c$ is a scaling parameter, $B_M = 1 - conf$ is the map's uncertainty, and $\overline{B_M}$ is the average uncertainty across the training dataset.

Finally, each map $M_i$ is binarized to create the refined segmentation mask $G_i$. For every pixel in $M_i$: if its intensity is greater than or equal to the calculated adaptive threshold ($th$), the corresponding pixel in $G_i$ is set to 1 (foreground); otherwise, it's set to 0 (background). This adaptive binarization ensures robust and accurate results.

**Region-Level Feature Refinement** The initial masks $G = \{G_i\}_{i=1}^{B}$ may lack holistic semantic context. To address this (as shown in Stage 3 of Fig. 2), we introduce a region-level feature refinement algorithm. This method discards disconnected regions whose feature representations are dissimilar to the overall common foreground.

The refinement begins by establishing a group-level foreground prototype ($F_G \in R^{1\times C}$), which is the average ViT patch embedding corresponding to the aggregated masks $G$:

$$F_G = \frac{1}{B}\sum_{n=1}^{B}\frac{1}{Area(G_i)}\sum F^i_{enhanced} \odot G_i. \tag{7}$$

For each image $I_i$ and its corresponding mask $G_i$:

1) Connected component labeling identifies individual sub-masks $\{L_{i,j}\}_{j=1}^{K_i}$, where $K_i$ is the number of components in $G_i$.

2) An all-zero refined mask, $R_i$, is initialized with the same dimensions as $G_i$.

3) For each sub-mask $L_{i,j}$, its feature embedding $F_{G_{i,j}}$ is computed by averaging the ViT patch embeddings within it:

$$F_{G_{i,j}} = \frac{1}{Area(L_{i,j})}\sum F^i_{enhanced} \odot L_{i,j}. \tag{8}$$

4) The cosine similarity $sim = \cos(F_G, F_{G_{i,j}})$ is calculated.

5) If $sim$ exceeds a threshold $sim_{th}$, $L_{i,j}$ is integrated into $R_i$.

The output is a set of refined segmentation masks $R = \{R_i\}_{i=1}^{N}$, with non-co-salient regions effectively removed.

Finally, to further enhance spatial coherence and boundary precision, we apply Dense Conditional Random Fields (DenseCRFs) to obtain the final $\hat{Y}$. This post-processing step ensures smooth regions and sharply defined object boundaries in the final segmentation masks.

The objective function for the predicted co-salient map includes the BCE loss and IoU loss, which are defined as follows:

$$L_{bce} = -\frac{1}{N}\sum\left[Ylog(\hat{Y}) + (1-Y)log(1-\hat{Y})\right], \tag{9}$$

$$L_{iou} = 1 - \frac{1}{N}\sum\frac{\hat{Y}\cap Y}{\hat{Y}\cup Y}, \tag{10}$$

where $\hat{Y}$ denotes for predicted co-salient maps and $Y$ denotes for pseudo-label. Then, the final objective function is: $L_{final} = L_{iou} + L_{bce}$.

# 3 Experiments

## 3.1 Setup

**Datasets and evaluation metrics:** We use two datasets as our training sets, namely DUTS-class [32] and COCO9k [17]. For testing, we employ three datasets: CoCA [32], CoSOD3k [31], and Cosal2015 [13]. CoCA: Frequently cited as a challenging dataset with 1,295 images across 80 categories. CoSOD3k: Described as the largest CoSOD evaluation dataset with 3,316 images in 160 groups. Cosal2015: A widely used dataset with 2,015 images in 50 groups. We evaluate the model using four metrics: Mean Absolute Error ($MAE$) [8], maximum F-measure ($F_\beta^{max}$) [1], E-measure ($E_\phi^{max}$) [12], S-measure ($S_\alpha$) [11]. The evaluation tool can be accessed at eval-co-sod [12, 33].

## 3.2 Implementation Details

For feature extraction, we employ pre-trained Vision Transformer (ViT) [10] models from DINO [4] and DINOv2 [19]. Specifically, we utilize a DINO ViT-B model with a patch size of 8, and a DINOv2 ViT-B-reg model with a patch size of 14. During training, the batch size is set to 24, and the model is trained for 100 epochs. We use the Adam optimizer with an initial learning rate of 0.0001. All experiments are conducted on an NVIDIA GeForce RTX 4070 Ti SUPER GPU, with a total training duration of approximately 12 hours.

**Table 1.** The comparison results of our UMS-CoSOD method with other state-of-the-art fully supervised and unsupervised methods, where **red** indicates the optimal and **blue** indicates the suboptimal. ↑ means that larger is better and ↓ denotes that smaller is better.

| Methods | Pub. | CoCA | | | | CoSOD3k | | | | CoSal2015 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $MAE\downarrow$ | $F_\beta^{max}\uparrow$ | $E_\phi^{max}\uparrow$ | $S_\alpha\uparrow$ | $MAE\downarrow$ | $F_\beta^{max}\uparrow$ | $E_\phi^{max}\uparrow$ | $S_\alpha\uparrow$ | $MAE\downarrow$ | $F_\beta^{max}\uparrow$ | $E_\phi^{max}\uparrow$ | $S_\alpha\uparrow$ |
| DCFM [29] | CVPR(2022) | 0.085 | 0.598 | 0.783 | 0.710 | 0.067 | 0.805 | 0.874 | 0.810 | 0.077 | 0.828 | 0.879 | 0.827 |
| CoGEM [26] | CVPR(2023) | 0.095 | 0.599 | 0.808 | 0.726 | 0.061 | 0.829 | **0.911** | 0.853 | 0.053 | 0.882 | 0.933 | 0.885 |
| DMT [16] | CVPR(2023) | 0.108 | 0.619 | 0.800 | 0.725 | 0.063 | 0.835 | 0.895 | 0.851 | 0.045 | 0.905 | 0.936 | 0.897 |
| GCoNet+ [34] | TPAMI(2023) | 0.081 | 0.637 | 0.814 | 0.738 | 0.062 | 0.834 | 0.901 | 0.843 | 0.056 | 0.891 | 0.924 | 0.881 |
| CONDA [15] | ECCV(2024) | 0.092 | **0.675** | **0.825** | **0.757** | 0.060 | **0.844** | 0.899 | **0.857** | 0.042 | 0.912 | 0.940 | 0.904 |
| Sg-CoSOD [23] | TCSVT(2024) | 0.185 | 0.400 | 0.666 | 0.591 | 0.124 | 0.684 | 0.786 | 0.748 | 0.117 | 0.735 | 0.809 | 0.779 |
| ESCF [7] | KBS(2025) | 0.084 | 0.598 | 0.802 | 0.717 | 0.057 | 0.838 | 0.905 | 0.851 | 0.040 | 0.902 | 0.942 | 0.894 |
| MGCNet [14] | KBS(2025) | 0.092 | 0.600 | 0.786 | 0.071 | 0.057 | 0.827 | 0.897 | 0.837 | 0.048 | 0.887 | 0.923 | 0.877 |
| TRNet [9] | TCSVT(2025) | 0.114 | 0.622 | 0.812 | 0.715 | 0.070 | 0.828 | 0.891 | 0.844 | 0.058 | 0.894 | 0.925 | 0.885 |
| CFPAM [18] | ICASSP(2025) | **0.074** | 0.658 | 0.814 | 0.753 | **0.050** | 0.828 | 0.892 | 0.850 | **0.035** | **0.917** | **0.944** | **0.906** |
| VCP [24] | CVPR(2025) | **0.069** | **0.680** | **0.829** | **0.774** | **0.049** | **0.868** | **0.918** | **0.874** | **0.037** | **0.920** | **0.944** | **0.911** |
| TokenCut [25] | CVPR(2022) | 0.167 | 0.467 | 0.704 | 0.627 | 0.151 | 0.720 | 0.811 | 0.744 | 0.139 | 0.805 | 0.857 | 0.793 |
| DVFDVD [2] | ECCVW(2022) | 0.223 | 0.422 | 0.592 | 0.581 | 0.104 | 0.722 | 0.819 | 0.773 | 0.092 | 0.777 | 0.842 | 0.809 |
| SegSwap [20] | CVPR(2022) | 0.165 | 0.422 | 0.666 | 0.567 | 0.177 | 0.560 | 0.705 | 0.608 | 0.178 | 0.618 | 0.720 | 0.632 |
| ZS-CoSOD [27] | ICASSP(2024) | 0.115 | 0.549 | - | 0.667 | 0.117 | 0.691 | - | 0.723 | 0.101 | 0.799 | - | 0.785 |
| US-CoSOD [5] | WACV(2024) | 0.116 | 0.546 | 0.743 | 0.672 | 0.076 | 0.779 | 0.861 | 0.801 | 0.070 | 0.845 | 0.886 | 0.840 |
| SCoSPARC [6] | ECCV(2024) | **0.092** | **0.614** | **0.782** | **0.711** | **0.064** | **0.827** | **0.889** | **0.823** | **0.062** | **0.869** | **0.905** | **0.851** |
| UMS(Ours) | | **0.082** | **0.621** | **0.794** | **0.725** | **0.051** | **0.863** | **0.918** | **0.859** | **0.057** | **0.881** | **0.924** | **0.857** |

**Table 2.** Comparison of our model with fully supervised CoSOD models trained with varying amounts of labeled data.

| | | CoCA | | | | CoSOD3k | | | | CoSal2015 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | Label | $MAE \downarrow$ | $F_\beta^{max} \uparrow$ | $E_\phi^{max} \uparrow$ | $S_\alpha \uparrow$ | $MAE \downarrow$ | $F_\beta^{max} \uparrow$ | $E_\phi^{max} \uparrow$ | $S_\alpha \uparrow$ | $MAE \downarrow$ | $F_\beta^{max} \uparrow$ | $E_\phi^{max} \uparrow$ | $S_\alpha \uparrow$ |
| DCFM | 50% | 0.107 | 0.562 | 0.752 | 0.686 | 0.090 | 0.767 | 0.851 | 0.774 | 0.102 | 0.799 | 0.844 | 0.802 |
| MCCL | 50% | 0.129 | 0.553 | 0.773 | 0.692 | 0.088 | 0.812 | 0.882 | 0.832 | 0.076 | 0.859 | 0.893 | 0.870 |
| GCoNet+ | 50% | 0.133 | 0.534 | 0.753 | 0.661 | 0.074 | 0.842 | 0.889 | 0.842 | 0.079 | 0.783 | 0.865 | 0.808 |
| CONDA | 50% | 0.122 | 0.626 | 0.780 | 0.715 | 0.089 | 0.803 | 0.857 | 0.827 | 0.070 | 0.872 | 0.898 | 0.864 |
| DCFM | 75% | **0.102** | 0.582 | 0.766 | 0.691 | 0.079 | 0.786 | 0.857 | 0.799 | 0.093 | 0.810 | 0.861 | 0.813 |
| MCCL | 75% | 0.116 | 0.580 | 0.776 | 0.696 | 0.081 | 0.825 | 0.892 | 0.842 | 0.065 | 0.881 | 0.913 | **0.878** |
| GCoNet+ | 75% | 0.113 | 0.547 | 0.759 | 0.682 | 0.087 | 0.842 | 0.892 | 0.846 | 0.071 | 0.804 | 0.876 | 0.823 |
| CONDA | 75% | 0.105 | **0.647** | **0.796** | **0.733** | **0.071** | **0.856** | **0.906** | **0.849** | **0.059** | **0.891** | **0.914** | **0.880** |
| Ours | 0% | **0.082** | **0.621** | **0.794** | **0.725** | **0.051** | **0.863** | **0.918** | **0.859** | **0.057** | **0.881** | **0.924** | 0.857 |

### 3.3 Quantitative Evaluation

**Comparison with State-of-the-Art Methods.** As presented in Table 1, we compare our proposed UMS-CoSOD method with state-of-the-art approaches in CoSOD, encompassing both fully supervised and unsupervised techniques. The table is divided by a double horizontal line, with fully supervised methods listed above and unsupervised methods below. Our method demonstrates a clear superiority over the leading unsupervised approaches. Specifically, on the largest CoSOD dataset, CoSOD3k, our method achieves significant improvements across all four evaluation metrics: a 19.8% reduction in $MAE$, a 4.4% increase in $F_\beta^{max}$, a 3.3% gain in $E_\phi^{max}$, and a 4.4% enhancement in $S_\alpha$ compared to the best unsupervised baselines.

These substantial improvements favorably validate the effectiveness of our approach. The notable performance gains can be attributed to the innovative design of our framework, particularly the CM-SPLG and the Unified Multi-Granularity Feature Learning strategy. By integrating instance segmentation masks from HASSOD and self-attention maps from DINO, our method generates high-quality pseudo-ground truth maps that accurately capture co-salient objects. Furthermore, the AFMM and CFMM enhance feature representation by adaptively fusing and refining features from multiple self-supervised backbones, leading to superior detection of co-salient objects across diverse and challenging scenarios.

When compared to the latest fully supervised methods, our approach exhibits remarkable competitiveness. On the CoSOD3k dataset, UMS-CoSOD outperforms several established fully supervised methods, including TRNet, MGCNet, ESCF, Sg-CoSOD, CONDA, GCoNet+, and DMT, with performance metrics closely approaching those of the state-of-the-art fully supervised method, VCP. This highlights the robustness of our self-supervised framework, which rivals fully supervised techniques despite not relying on extensive manual annotations.

Table 2 further illustrates the comparison between our method and four fully supervised approaches—DCFM, MCCL, GCoNet+, and CONDA—under conditions with limited annotated datasets. We randomly sampled 50% and 75% of the labeled datasets from each category group in the training set to train these four methods. The results indicate that UMS-CoSOD consistently achieves ei-

ther the best or second-best performance across these scenarios. Notably, on the CoSOD3k dataset, our method surpasses all four fully supervised methods across all metrics. This superior performance underscores the capability of our approach to effectively leverage limited annotated data, achieving results comparable to or better than fully supervised counterparts. The success in this context is driven by the robust pseudo-label generation process of CM-SPLG, which provides high-quality training signals, and the multi-stage refinement pipeline—including the FRAM and RRBE—which ensures generalization and precision even with scarce annotations. This makes our method particularly advantageous in practical settings where labeled data is limited or costly to obtain.

**Table 3.** The results of the ablation study.

| | | | | CoCA | | | | CoSOD3k | | | | CoSal2015 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AFMM | CFMM | FRAM | RRBE | $MAE \downarrow$ | $F_\beta^{max} \uparrow$ | $E_\phi^{max} \uparrow$ | $S_\alpha \uparrow$ | $MAE \downarrow$ | $F_\beta^{max} \uparrow$ | $E_\phi^{max} \uparrow$ | $S_\alpha \uparrow$ | $MAE \downarrow$ | $F_\beta^{max} \uparrow$ | $E_\phi^{max} \uparrow$ | $S_\alpha \uparrow$ |
| | | | | 0.122 | 0.144 | 0.323 | 0.459 | 0.189 | 0.240 | 0.369 | 0.454 | 0.223 | 0.280 | 0.352 | 0.436 |
| ✓ | ✓ | | | 0.140 | 0.505 | 0.703 | 0.634 | 0.144 | 0.610 | 0.744 | 0.672 | 0.150 | 0.628 | 0.737 | 0.672 |
| ✓ | ✓ | ✓ | | **0.093** | **0.609** | **0.780** | **0.708** | 0.063 | **0.827** | **0.889** | **0.825** | **0.062** | **0.867** | **0.908** | **0.854** |
| ✓ | ✓ | | ✓ | 0.106 | 0.560 | 0.756 | 0.676 | **0.060** | 0.798 | 0.877 | 0.808 | 0.070 | 0.836 | 0.892 | 0.831 |
| ✓ | ✓ | ✓ | ✓ | **0.082** | **0.621** | **0.794** | **0.725** | **0.051** | **0.863** | **0.918** | **0.859** | **0.057** | **0.881** | **0.924** | **0.857** |


**Fig. 5.** Our UMS-CoSOD is qualitatively compared with other methods on the CoCA, Cosal2015, and CoSOD3k datasets, where the blue box indicates unsupervised methods and the red box indicates fully supervised methods.

**Ablation Studies.** To rigorously assess the contribution of each component within our UMS-CoSOD framework, we conducted extensive ablation studies on the CoCA, CoSOD3k, and CoSal2015 datasets, with results detailed in Table 3. Our initial baseline (Row 1) involved a simple fusion of output features from

DINO and DINOv2, where their respective features were element-wise added with a 0.5 weighting. Introducing the AFMM and the CFMM (Row 2) on top of this baseline led to notable improvements in co-salient feature localization, evidenced by substantial gains in $F_\beta^{max}$ (e.g., from 0.240 to 0.610 on CoSOD3k) and $S_\alpha$ (e.g., from 0.454 to 0.672 on CoSOD3k), demonstrating their effectiveness in initial feature aggregation and inter-image feature mining. However, the most striking performance leap was observed with the integration of the Feature Refinement and Amplification Module (FRAM) (Row 3, highlighted in blue). The inclusion of FRAM dramatically enhanced the results across all metrics and datasets; for instance, on CoSal2015, $MAE$ plummeted from 0.150 to an impressive 0.062, and $F_\beta^{max}$ surged from 0.628 to 0.867. Similarly, on CoSOD3k, FRAM boosted $F_\beta^{max}$ from 0.610 to 0.827 and $E_\phi^{max}$ from 0.744 to 0.889, clearly underscoring its critical role in refining features and decisively amplifying salient information essential for accurate co-saliency detection. The RRBE stage further validated its utility, both independently improving results over the AFMM+CFMM configuration (Row 4 vs. Row 2, e.g., reducing $MAE$ on CoSOD3k from 0.144 to 0.060) and providing the final polish in the complete model. Ultimately, our full model (Row 5, highlighted in red), incorporating all proposed modules, achieved the overall best performance (e.g., $MAE$ of 0.051 and $F_\beta^{max}$ of 0.863 on CoSOD3k; $MAE$ of 0.057 and $F_\beta^{max}$ of 0.881 on CoSal2015), confirming the effective, synergistic design where modules like AFMM and CFMM lay a solid foundation, FRAM provides the crucial amplification and refinement of salient features, and RRBE ensures precise final outputs.

### 3.4 Qualitative Evaluation

Fig. 5 presents a qualitative comparison of UMS-CoSOD against SCoSPARC (self-supervised), US-CoSOD, ZS-CoSOD (unsupervised), and DMT, UFGS, GCoNet+ (fully supervised) on challenging image groups from CoCA, CoSOD3k, and CoSal2015.

As shown in Fig. 5, UMS-CoSOD consistently demonstrates superior performance, especially in complex multi-object scenarios. For example, in the CoCA Camera group with distractors, our method yields outstanding results, often surpassing fully supervised methods. This is largely attributed to our AFMM, which intelligently fuses robust DINO and DINOv2 features with adaptive weighting and cross-modal attention, suppressing background noise and highlighting co-salient cues.

Our model also delivers excellent segmentation with diverse object appearances and intricate backgrounds, as seen in CoSOD3k's Remote Control (shape variations) and CoSal2015's Goldenfish (color patterns). UMS-CoSOD's proficiency in these scenarios results from the synergistic interplay of its modules:

CFMM excels at discerning invariant features despite intra-group diversity, effectively capturing the shared essence of common objects. FRAM enhances discriminative power via adaptive cross-scale attention, refining co-saliency signals and suppressing background noise for sharp boundaries. RRBE produces high-fidelity final masks through confidence-based adaptive thresholding, region-level

refinement (eliminating non-co-salient components), and DenseCRFs, ensuring accurate, visually appealing, and semantically consistent output.

## 4 Conclusion

This paper proposes UMS-CoSOD, a novel self-supervised framework for co-salient object detection that eliminates manual annotations. By integrating Cross-Modal Self-Supervised Pseudo-Label Generation (CM-SPLG) with Feature Learning and Saliency Prediction (featuring DINO/DINOv2 backbones, AFMM, CFMM, and FRAM), and Result Refinement and Boundary Enhancement (RRBE), our approach achieves robust co-saliency detection. UMS-CoSOD outperforms state-of-the-art unsupervised methods on three benchmark datasets and remains competitive with fully supervised techniques. Its annotation-free design offers high practicality for real-world applications.

**Acknowledgement** This work is supported by the Qingdao Natural Science Foundation under Grant 24-4-4-zrjj-90-jch, and the Shandong Provincial Natural Science Foundation under Grant ZR2024QF267, the Major Innovative Project of Science, Education and Industry Integration Pilot Project, Qilu University of Technology (Shandong Academy of Sciences) under Grant 2024ZDZX08, and the Pairing Plan Project of Faculty of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences) under Grant Nos. 2024JDJH07, and the National Natural Science Foundation of China under Grant 62502287.

## References

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 1597–1604. IEEE (2009)
2. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep vit features as dense visual descriptors. arXiv preprint arXiv:2112.05814 **2**(3), 4 (2021)
3. Cao, S., Joshi, D., Gui, L., Wang, Y.X.: Hassod: Hierarchical adaptive self-supervised object detection. Advances in Neural Information Processing Systems **36**, 59337–59359 (2023)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
5. Chakraborty, S., Naha, S., Bastan, M., Samaras, D., et al.: Unsupervised and semi-supervised co-salient object detection via segmentation frequency statistics. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 332–342 (2024)
6. Chakraborty, S., Samaras, D.: Self-supervised co-salient object detection via feature correspondences at multiple scales. In: European Conference on Computer Vision. pp. 231–250. Springer (2024)

7. Chen, Q., Wang, Q., Fang, X., Sun, Q., Zhu, J.: Edge and semantic collaboration framework with cross coordination attention for co-saliency detection. Knowledge-Based Systems **318**, 113513 (2025)
8. Cheng, M.M., Warrell, J., Lin, W.Y., Zheng, S., Vineet, V., Crook, N.: Efficient salient region detection with soft image abstraction. In: Proceedings of the IEEE International Conference on Computer vision. pp. 1529–1536 (2013)
9. Cong, R., Yang, N., Liu, H., Zhang, D., Huang, Q., Kwong, S., Zhang, W.: Trnet: Two-tier recursion network for co-salient object detection. IEEE Transactions on Circuits and Systems for Video Technology (2025)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision. pp. 4548–4557 (2017)
12. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421 (2018)
13. Fan, D.P., Lin, Z., Ji, G.P., Zhang, D., Fu, H., Cheng, M.M.: Taking a deeper look at co-salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2919–2929 (2020)
14. Fang, X., Wang, X., Zhu, J., Chen, Q., Chen, Z.: Mgcnet: Multiple group-wise correlation network with hierarchical contrastive learning for co-salient object detection. Knowledge-Based Systems **309**, 112852 (2025)
15. Li, L., Liu, N., Zhang, D., Li, Z., Khan, S., Anwer, R., Cholakkal, H., Han, J., Khan, F.S.: Conda: Condensed deep association learning for co-salient object detection. In: European Conference on Computer Vision. pp. 287–303. Springer (2024)
16. Li, L., Xie, H., Liu, N., Zhang, D., Anwer, R.M., Cholakkal, H., Han, J.: Advanced discriminative co-saliency and background mining transformer for co-salient object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
18. Lv, G., Yuan, M., Zhang, Z., Zhang, Z., Ding, Z., Ma, G.: Comprehensive feature processing based on attention mechanism for co-salient object detection. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
19. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
20. Shen, X., Efros, A.A., Joulin, A., Aubry, M.: Learning co-segmentation by segment swapping for retrieval and discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5082–5092 (2022)
21. Vanyan, A., Barseghyan, A., Tamazyan, H., Huroyan, V., Khachatrian, H., Danelljan, M.: Analyzing local representations of self-supervised vision transformers. arXiv preprint arXiv:2401.00463 (2023)
22. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)

23. Wang, J., Yu, N., Zhang, Z., Han, Y.: Single-group generalized rgb and rgb-d co-salient object detection. IEEE Transactions on Circuits and Systems for Video Technology (2024)
24. Wang, J., Yu, N., Zhang, Z., Han, Y.: Visual consensus prompting for co-salient object detection. arXiv preprint arXiv:2504.14254 (2025)
25. Wang, Y., Shen, X., Hu, S.X., Yuan, Y., Crowley, J.L., Vaufreydaz, D.: Self-supervised transformers for unsupervised object discovery using normalized cut. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14543–14553 (2022)
26. Wu, Y., Song, H., Liu, B., Zhang, K., Liu, D.: Co-salient object detection with uncertainty-aware group exchange-masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19639–19648 (2023)
27. Xiao, H., Tang, L., Li, B., Luo, Z., Li, S.: Zero-shot co-salient object detection framework. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4010–4014. IEEE (2024)
28. Xu, P., Mu, Y.: Co-salient object detection with semantic-level consensus extraction and dispersion. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 2744–2755 (2023)
29. Yu, S., Xiao, J., Zhang, B., Lim, E.G.: Democracy does matter: Comprehensive feature mining for co-salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 979–988 (2022)
30. Yu, Y., Zhang, Y., Cheng, Z., Song, Z., Tang, C.: Mca: Multidimensional collaborative attention in deep convolutional neural networks for image recognition. Engineering Applications of Artificial Intelligence **126**, 107079 (2023)
31. Zhang, D., Han, J., Li, C., Wang, J.: Co-saliency detection via looking deep and wide. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2994–3002 (2015)
32. Zhang, Z., Jin, W., Xu, J., Cheng, M.M.: Gradient-induced co-saliency detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. pp. 455–472. Springer (2020)
33. Zhang, Z., Jin, W., Xu, J., Cheng, M.M.: Gradient-induced co-saliency detection. In: European Conference on Computer Vision (ECCV) (2020)
34. Zheng, P., Fu, H., Fan, D.P., Fan, Q., Qin, J., Tai, Y.W., Tang, C.K., Van Gool, L.: Gconet+: A stronger group collaborative co-salient object detector. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 10929–10946 (2023)
35. Zhou, C., Wang, Z., Zhou, Y., Pan, C.: Sdnet: A simple and efficient salient object detection decoder with only 60k parameters. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 566–580. Springer (2024)