Unsupervised Salient Object Detection with Pseudo-Labels Refinement

Yanfeng Zheng[1], Pengjie Wang[1,2], Hao Liu[1], and Xiaosong Yang[2⋆]

[1] Dalian Minzu University, Dalian 116650, China
Zhengyanfeng1998@163.com
pengjiewang@gmail.com
202412054063@stu.dlnu.edu.cn
[2] Bournemouth University, Fern Barrow, Poole, Dorset, BH12 5BB, United Kingdom
xyang@bournemouth.ac.uk

Abstract. In Salient Object Detection(SOD), most methods rely on manually annotated labels, which are costly. As a result, unsupervised methods have gained significant attention. Existing methods often generate noisy pseudo-labels using traditional techniques, which can affect model performance. To address this, we propose an unsupervised method for RGB image salient object detection that generates high-quality pseudo-labels without manual annotation and uses them to train the detection model.The method generates initial pseudo-labels and improves their quality by introducing contrastive learning pre-trained weights and a pseudo-label self-updating strategy. Additionally, we design a detection network with a Multi-Feature Aggregation (MFA) module and a Context Feature Interaction (CFI) module to enhance the model's ability to detect salient objects in complex scenarios. The model we proposed, trained with our pseudo-labels, shows significant improvement on USOD and achieves excellent scores on public benchmarks.

Keywords: Unsupervised · Salient Object Detection · Contrastive Learning · Pseudo-Labels.

1 Introduction

The development of deep learning has significantly advanced salient object detection, with fully-supervised methods achieving notable breakthroughs. However, these methods are highly dependent on large-scale, accurately labeled data. To reduce the annotation burden, weakly-supervised methods have emerged, such as class labels [1] text descriptions [2],bounding boxes [3], scribbles [4] and point annotations [5]. Despite progress, human annotation is still required. Unsupervised methods aim to eliminate the need for human annotations altogether, offering better applicability in real-world scenarios where labeled data is scarce. A key challenge for unsupervised methods is generating high-quality pseudo-labels through image modeling, which is essential for training effective models.

⋆ *Corresponding author

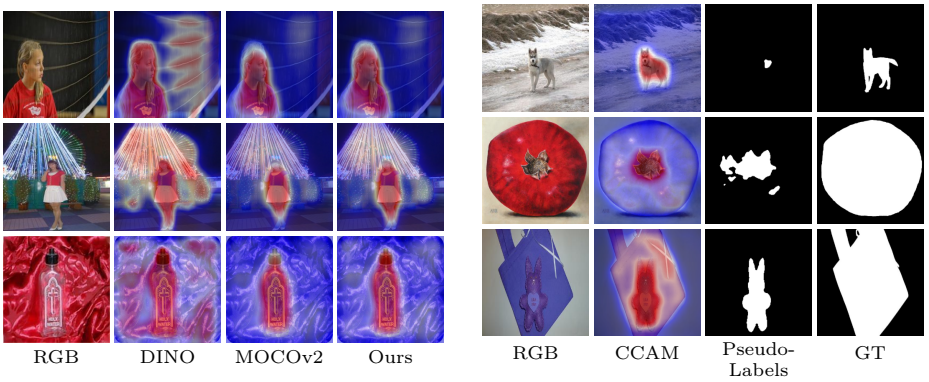

Fig. 1. (a) Visualisation of class-agnostic activation maps for different pre-trained weights. (b) Incorrect pseudo-labeling results.

Before the rise of deep learning, unsupervised methods mainly relied on handcrafted features like color contrast to identify salient regions, but these methods struggled in complex scenes. Today, most unsupervised methods generate initial pseudo-labels using traditional techniques and refine them with various strategies. However, traditional methods often produce low-quality pseudo-labels, limiting detection performance. Researchers are exploring advanced algorithms to improve pseudo-label accuracy and overall detection. Few methods use deep learning for pseudo-label generation, but Zhou et al. [6] showed that pre-trained weights from convtrastive learning can provide supervision for salient object detection models, yielding impressive results. One such method, CCAM [7], uses unsupervised contrastive learning to identify foreground regions by contrasting foreground and background in different images. As shown in Figure 1, CCAM trained with MOCOv2 [8] weights achieves good foreground localization but incomplete coverage, while CCAM trained with DINO weights [9] provides full coverage but with redundancy. These issues affect the quality of the final pseudo-labels.

In generating category-agnostic activation maps and refining them with a dense conditional random field (DCRF) to produce pseudo-labels, several challenges arise, as shown in Figure 1. While activation maps highlight target regions, they often lack precise edges, and complex scenes present further refinement difficulties. Additionally, some activation regions may not be suitable for salient object detection, leading to inaccurate pseudo-labels. To address these issues, this paper proposes a two-stage model for salient object detection. The first stage generates pseudo-labels in two steps: enhancing the original CCAM using offline distillation for the initial pseudo-label network, and refining the labels with a self-updating strategy. The second stage focuses on salient object detection, where the model is primarily supervised by the generated pseudo-labels. Key components of this model include: 1) a multi-feature aggregation module

to enhance high-level features, and 2) a context feature interaction module for improved feature fusion, boosting detection performance.

Our main contributions can be summarized as follows:

(1) This work introduces an updated pseudo-label generation method, leveraging different pre-trained weights for complementary learning and a self-updating strategy to improve label quality.

(2) A salient object detection network is designed to boost detection performance, incorporating a multi-feature aggregation module and a context feature interaction module.

(3) Experiments on four common RGB image saliency detection datasets demonstrate that the proposed method performs comparably to current weakly-supervised and unsupervised approaches.

2 Related work

2.1 Fully-Supervised Method Salient Object Detection

The majority of Salient Object Detection (SOD) methods are rely on extensive pixel-level manual annotations as the foundation for training and optimization. Qin et al. [10] proposed the BASNet method, which incorporates boundary-aware mechanisms to enhance the accuracy of salient object detection by focusing on the boundaries of objects. Liu et al. [11] proposed a feature aggregation module structure based on the U-net structure, combining coarse-level and high-level information. Pang et al. [24] proposed a multi-scale interactive network that uses multi-scale features and interactive mechanisms to improve the accuracy of salient object detection. Xu et al. [13] proposed PA-KRN, a progressive architecture for salient object detection that first locates objects globally using a coarse module, then segments them locally with a fine module, and uses an attention-based sampler to highlight salient regions. Liang et al. [14] proposed ExPert, a parameter-efficient fine-tuning method for salient object detection that uses adapters and injectors in a frozen transformer encoder to incorporate external prompt features, achieving superior performance with fewer parameters.

2.2 Weakly-Supervised Method Salient Object Detection

The prevailing state-of-the-art techniques for salient object detection are heavily dependent on extensive datasets that require precise pixel-level manual annotations. The creation of such annotations is both time-consuming and labor-intensive. Consequently, weakly-supervised approaches are emerging as a prominent and increasingly favored research trajectory. Piao et al. [15] employed an iterative calibration strategy to mitigate the pseudo-labeling error within the network. Zhang et al. [16] conducted supervised training by annotating simple pairs of images with foreground and background labels. Piao et al. [17] introduced a multiple pseudo-label fusion framework that leverages richer information from multiple labels to diminish the impact of the algorithmic process. Gao et al. [18]

presented a point-supervised approach that initially acquires pseudo-labels via an adaptive masking algorithm and subsequently generates the final prediction saliency maps through a Transformer-based network.

2.3 Unsupervised Method Salient Object Detection

In the field of salient object detection, weakly-supervised methods have played a significant role, but unsupervised methods have also garnered considerable attention. Unsupervised methods aim to detect salient objects without any explicit annotations. Nguyen et al. [19] proposed the DeepUSPS method, which uses self-supervision to leverage the input image itself as a natural supervisory signal for robust unsupervised saliency prediction. Yan et al. [20] introduced an uncertainty-aware pseudo-label learning approach for unsupervised domain adaptation in salient object detection, enabling the model to adapt to the target domain without labeled data in that domain. Wang et al. [21] proposed a method for deep unsupervised saliency detection that mines multi-source uncertainty to select reliable labels from multiple noisy labels, thereby improving the performance of unsupervised saliency detection. Zhou et al. [6] introduced a method called "Activation to Saliency", which forms high-quality labels for unsupervised salient object detection by leveraging activation information, leading to better detection results. Zhou et al. [22] proposed a texture-guided saliency distilling method by matching textures around the predicted boundaries for unsupervised salient object detection.

3 Method

The unsupervised saliency object detection process discussed in this paper mainly consists of two key stages: the first is the pseudo-label generation stage, where pseudo-labels are generated based on RGB images; the second is the saliency object detection stage, which differs from fully-supervised methods in that it uses the pseudo-labels generated in the first stage for learning and supervision. In this section, we will first describe the method for generating pseudo-labels, and then introduce the two core modules that constitute the saliency object detection network, namely the Multi-Feature Aggregation module (MFA) and the Contextual Feature Interaction Module (CFI).

3.1 Pseudo-label generation model

This study proposes a novel method for generating pseudo-labels using class-agnostic activation maps, which automatically identify and locate salient objects. Instead of directly using the CCAM method, the network is enhanced with different pre-trained weights. A CCAM model trained with DINO pre-trained weights serves as an auxiliary supervision signal, providing additional guidance to improve training and combine the strengths of both weight sets.

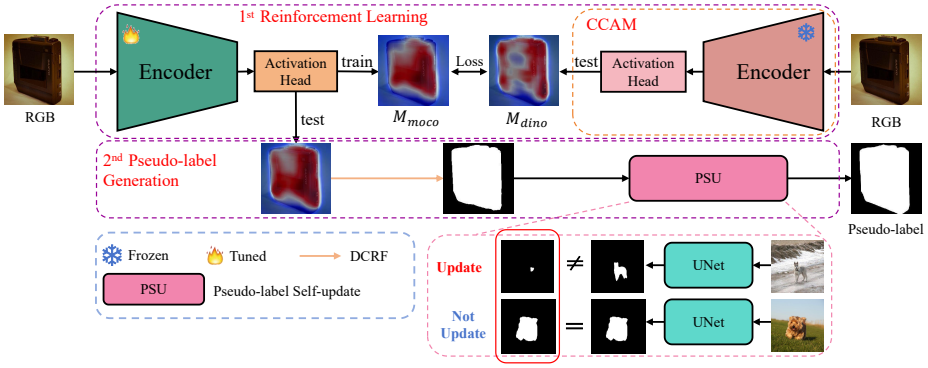

Fig. 2. Pseudo-label generation method structure

As shown in the upper part of Figure 2, in the specific implementation, Resnet-50 is used as the encoder of the backbone network. An RGB image is input, and after being processed by the encoder of the backbone network, four sets of feature maps F_1, F_2, F_3, and F_4 are obtained. This process can be represented as:

$$F_1, F_2, F_3, F_4 = Encoder(I_m) \tag{1}$$

Here, I_m represents the input RGB image, and Encoder represents the encoder. Then, the feature maps F_3 and F_4 are concatenated along the channel dimension and then processed through the CBS operation to generate the class-agnostic activation map M_{moco}, This process can be represented as:

$$M_{moco} = CBS(Contact(F_3, F_4)) \tag{2}$$

Here, $Concat()$ denotes the concatenation operation along the channel dimension, and CBS represents a sequence of operations including a 3×3 convolution, BatchNorm, and a Sigmoid activation function. Additionally, based on the aforementioned process, the encoder is pre-trained using DINO pre-trained weights to generate a class-agnostic activation map represented as M_{dino}.

$$\mathcal{L} = \mathcal{L}_{\text{POS}} + \mathcal{L}_{\text{NEG}} + \alpha\mathcal{L}_{\text{SSIM}} + \beta\mathcal{L}_{\text{IoU}} \tag{3}$$

Here, $\mathcal{L}_{\text{POS}}$ and $\mathcal{L}_{\text{NEG}}$ are the original CCAM losses, $\mathcal{L}_{\text{SSIM}}$ is the structural similarity loss, and $\mathcal{L}_{\text{IOU}}$ is the intersection over union loss. The values of α and β are set to 0.2.

After generating the final class-agnostic activation maps using the aforementioned strategy, Dense Conditional Random Fields (DCRF) are further employed to process these activation maps to generate the initial pseudo-labels Y_{PL}. This process aims to refine the saliency maps from the original activation maps, providing more accurate labels for subsequent training. However, although DCRF can improve the quality of the labels to some extent, the pseudo-labels still have imperfections in detail, as shown in the first and second columns of the third row in Figure 1. Due to the characteristics of the class-agnostic activation maps,

some activated regions may not be entirely suitable for the task of salient object detection, as shown in the third and fourth columns of the third row in Figure 1. These incomplete or incorrect refinements, if used as the basis for long-term network training, may lead the model to learn these inaccurate pieces of information, ultimately affecting the detection performance of the network. Despite the

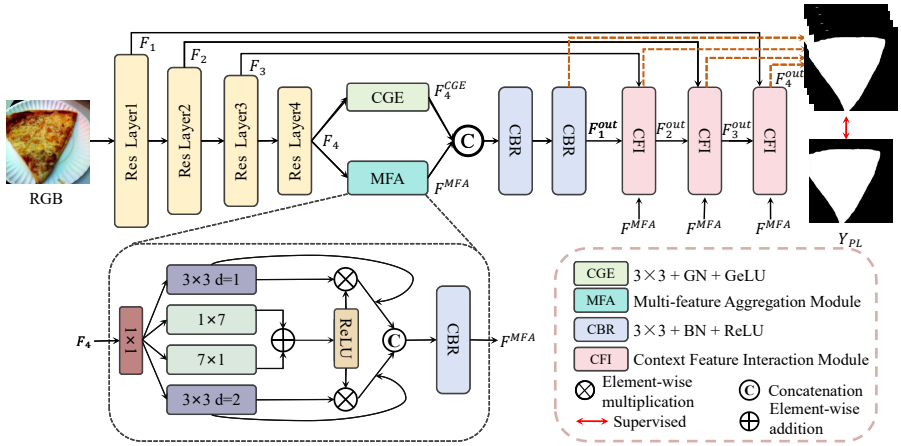

Fig. 3. The structure of the salient object detection network

potential inaccuracies in the pseudo-labels, network training remains an iterative learning and optimization process. Even with imperfect labels, they still guide the salient object detection network towards the correct targets, providing a generally valid learning direction. This demonstrates that the network can learn effective saliency information by capturing statistical patterns in large datasets, even with imprecise labels. In the early stages of training, the network is highly sensitive to the saliency information in the pseudo-labels, highlighting the importance of effective pseudo-label updating strategies. A well-designed updating strategy enhances the network's ability to capture saliency features, improving detection performance. Based on this, we propose a pseudo-label self-updating algorithm, as shown in the lower part of Figure 2. Specifically, the generated pseudo-labels Y_{PL} are used to train a simple U-shaped network, and the saliency map $Y_{PL}^{'}$ produced by the network is used to update the pseudo-labels. In the early stages, the model can more accurately identify and correct errors in the pseudo-labels, and iteratively updating them improves both their accuracy and detail, ultimately enhancing the detection performance.

In this algorithm, the pseudo-labels are self-updated using different evaluation criteria at different training stages to improve the model's performance. Specifically, in the 2nd to 5th rounds of training, the algorithm uses the intersection over union (IoU) to measure the similarity between the model's current predictions and the previous pseudo-labels.If the result is below the threshold, the pseudo-labels are updated using the current model predictions. In the later

stages of training, the pseudo-labels are updated using the Structure Similarity Index Measure (SSIM) [34] as the update criterion.

Here, the threshold is initially set to 0.9 for each evaluation criterion, and starting from the second epoch it is continuously updated during training, increasing by 0.1 each epoch over a total of 10 epochs. By dynamically adjusting the update strategy during training, the pseudo-labels are continuously refined, thereby enhancing the model's understanding of the data and the accuracy of its predictions.

3.2 Unsupervised Salient Object Detection with Pseudo-labels

To better enhance the performance of salient object detection, this paper designs a salient object detection model that uses Resnet-50 as the backbone network for feature extraction. An input RGB image is processed through the backbone network to obtain four features, namely F_1, F_2, F_3, and F_4, which are used as inputs for the multi-feature aggregation module and the context feature interaction module. The overall architecture is shown in Figure 3.

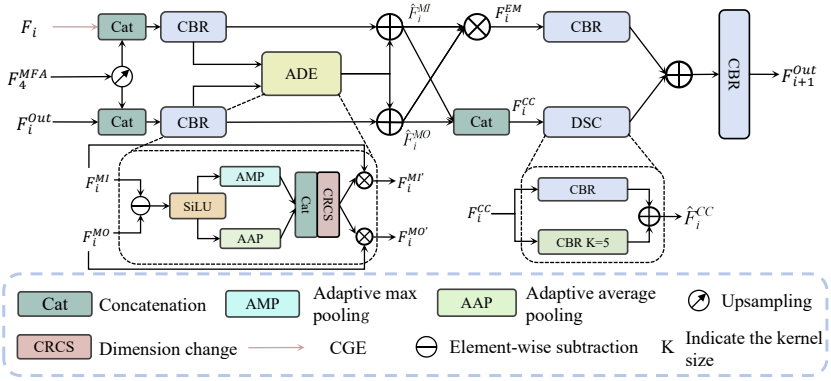

Fig. 4. Contextual Feature Interaction Module (CFI)

Multi-Feature Aggregation Module In deep learning tasks, the shallow layers of a network extract low-level features, while higher convolutional layers extract more advanced features. Among these, high-level semantic features are crucial as they provide a deep and abstract understanding of the image content. The abstract nature of these features enables them to effectively capture complex concepts and entities within the image, ensuring robustness against variations. By enhancing high-level semantic features, the model can more accurately understand and represent complex structures and abstract concepts within the image. Chen et al. [27] used dilated convolutions to expand the receptive field of convolutional layers, significantly improving the model's ability to recognize objects of different sizes without increasing the number of parameters or computational

burden. To this end, this paper designs a multi-feature aggregation module that primarily enhances the high-level feature F_4 from the encoder. By employing convolutional kernels of various sizes and shapes, the module enhances the feature representation and adapts to the processing needs of objects of different shapes. Specifically, as shown in the MFA (Multi-feature Aggregation) module in Figure 3, the input is F_4. First, a 1×1 convolution is applied to reduce the dimensionality of the feature, resulting in $F_4^{'}$. $F_4^{'}$ is then processed through 3×3 convolution operations with different dilation rates to obtain the features $\tilde{F}_4$ and$\bar{F}_4$. The process can be represented as:

$$\begin{aligned} \tilde{F}_4 &= \mathrm{Conv}_{d=1}(F_4') \\ \bar{F}_4 &= \mathrm{Conv}_{d=2}(F_4') \end{aligned} \tag{4}$$

Here, Conv denotes a convolution with a 3×3 kernel, and d represents the dilation rate. By combining vertical and horizontal convolution kernels, the model can more comprehensively capture spatial information in the image. Compared to using traditional 3×3 and 7×7 convolution kernels, this method not only reduces the number of parameters and the risk of overfitting but also increases the model's processing speed and efficiency. For this reason, $F_4^{'}$ is also processed through convolution kernels in different directions to obtain spatial information in the image and then passed through a ReLU layer to obtain F_{HW}. The process can be represented as:

$$F_{HW} = \mathrm{ReLU}(\mathrm{Conv}_H(F_4') \oplus \mathrm{Conv}_W(F_4')) \tag{5}$$

Here, $Conv_H$ denotes a vertical convolution with a 7×1 kernel, and $Conv_W$ denotes a horizontal convolution with a 1×7 kernel. The symbol $\otimes$ represents element-wise addition. To better integrate the features from dilated convolutions and the spatially enhanced features, the feature map F_{HW} is element-wise multiplied with the dilated features $\tilde{F}_4$ and $\bar{F}_4$ of different dilation rates. Additionally, skip connections are applied to each set of features to fuse the original features. This approach not only enhances the spatial representation but also maintains the integrity of the original features, thereby providing the network with a richer and more effective feature representation.

$$\begin{aligned} \tilde{F}_4 &= \tilde{F}_4 \copyright (F_{HW} \otimes \tilde{F}_4) \\ \bar{F}_4 &= \bar{F}_4 \copyright (F_{HW} \otimes \bar{F}_4) \end{aligned} \tag{6}$$

Here, $\oplus$ denotes element-wise multiplication. Finally,$\tilde{F}_4$ and $\bar{F}_4$ are concatenated and then passed through a CBR to obtain the feature F_{MFA}. The process can be represented as: Through the aforementioned operations, convolutional kernels of different shapes and sizes are effectively integrated, thereby significantly enhancing the feature representation capabilities. By expanding the receptive field, this method enables the network to learn richer spatial attributes, thereby deeply exploring and utilizing the complexity and diversity of image content. This enhances the high-level feature F_4 and provides richer and more effective input features for subsequent modules.

Context Feature Interaction Module In salient object detection, the U-shaped structure is commonly used for its strong performance. However, as high-level features pass upwards in this structure, their information density decreases, impacting detection capability [11]. To address this, we propose a Context Feature Interaction Module that enhances feature interaction across levels, mitigating the dilution of high-level features during transmission.

As shown in Figure 4, the inputs to this module are F_{MFA}, F_i^{Out}, and F_i, which originate from different stages of the model and each contain unique information and data representations. First, F_{FMA} is concatenated with F_i^{Out} and F_i respectively. Then, these concatenated features are processed through two separate CBRs to obtain two new features F_i^{MI} and F_i^{MO}. These features are then fed into the Adaptive Difference Enhancement Module (ADE).

The primary function of the ADE module is to calculate the differences between the two input features and process these difference features using the SiLU function to highlight important information and suppress less important information. Subsequently, the ADE module further processes these difference features through adaptive average pooling and adaptive max pooling operations. These two types of pooling operations extract features from different perspectives, and combining the pooled features helps to integrate their respective advantages. By applying these combined features to the original input features through element-wise multiplication, the expressive power of the input features is further enhanced. Additionally, skip connections are introduced to prevent information loss during the weighting process, resulting in $\hat{F}_i^{MI}$ and $\hat{F}_i^{MO}$. The process is as follows:

$$\begin{aligned} F_i^{MI'}, F_i^{MO'} &= ADE(F_i^{MI}, F_i^{MO}) \\ \hat{F}_i^{MI} &= F_i^{MI'} + F_i^{MI} \\ \hat{F}_i^{MO} &= F_i^{MO'} + F_i^{MO} \\ F_i^{CC} &= \mathrm{Cat}(\hat{F}_i^{MI}, \hat{F}_i^{MO}) \end{aligned} \tag{7}$$

In the feature interaction operation, $\hat{F}_i^{MI}$ and $\hat{F}_i^{MO}$ are element-wise multiplied to generate F_i^{EM}, which helps to capture and enhance the interactions and dependencies between the two features.

$$F_i^{EM} = \hat{F}_i^{MI} \otimes \hat{F}_i^{MO} \tag{8}$$

To enhance the representation capability of the feature $\hat{F}_i^{CC}$, a multi-scale convolutional kernel strategy is employed to capture different scale information from the input features. Specifically, convolutional kernels of different sizes are applied to $\hat{F}_i^{CC}$ to extract features at different scales, and these features are then element-wise added to obtain F_i^{CC}. The process can be represented as:

$$\hat{F}_i^{CC} = CBR(F_i^{CC}) + CBR_{k=5}(F_i^{CC}) \tag{9}$$

By integrating features from different scales, the expressiveness and adaptability of the features are further enhanced. Finally, to combine multiple feature representations, $\hat{F}_i^{CC}$ andF_i^{EM} are element-wise added and then processed through a

CBR operation to obtain the final output feature F_{i+1}^{Out} of the Context Feature Interaction Module. The process can be represented as:

$$F_{i+1}^{Out} = CBR(\hat{F}_i^{CC} + F_i^{EM}) \tag{10}$$

This paper replaces the traditional U-shaped structure's decoder with the Context Feature Interaction Module, which more effectively integrates feature information across different levels, particularly during upsampling and resolution restoration. This module combines deep semantic information with shallow detail, enhancing the model's ability to capture target details and improving overall feature representation. As a result, the model better incorporates both contextual and local information during decoding, boosting performance.

3.3 Loss Function

In this paper, a combined loss function is used for training, which includes the intersection over union loss ($\mathcal{L}_{IoU}$) and the local saliency coherence loss ($\mathcal{L}_{lsc}$) [25]. Additionally, this paper employs a deep supervision strategy, which introduces supervision signals at different network layers to further improve the model's performance. The formula for the total loss in this paper is as follows:

$$\mathcal{L} = \sum_{i=1}^{4} \left(\mathcal{L}_{IoU} \left(Y_i^{out}, Y_{pl} \right) + \mathcal{L}_{lsc} \right) \tag{11}$$

4 Experiments and results

4.1 Datasets

In the experiments of this paper, DUTS-TR [30], is used as the training dataset. The pixel-level pseudo-labels generated by the proposed method serve as supervision signals for network training. For testing, the method is evaluated on ECSSD [31], DUTS-TE [30], DUT-OMRON [32], and HKU-IS [33] datasets.

4.2 Experimental Details

Experiments were conducted on a NVIDIA GTX 3090 GPU using the PyTorch framework. The first stage's hyperparameters match those of CCAM, while the second stage uses a DINO pre-trained ResNet-50 as the backbone. Training images are resized to 256 $\times$ 256, with the Adam optimizer and a batch size of 32. The model trains for 15 epochs, starting with a learning rate of 1e-4, which decays by 10% every 5 epochs.

4.3 Evaluation Metrics

This paper employs three commonly used evaluation metrics in salient object detection, to assess the performance of different models. These include the F-measure (F_β) [28], Mean Absolute Error (MAE) [29], E-measure [26].

Table 1. Quantitative comparisons on four datasets

Method	Sup	DUTS-TE			HKU-IS			ECSSD			DUT-OMRON		
		$MAE \downarrow$	$E_m \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$E_m \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$E_m \uparrow$	$F_\beta \uparrow$	$MAE \downarrow$	$E_m \uparrow$	$F_\beta \uparrow$
RBD [10]	T	0.162	0.664	0.428	0.176	0.716	0.54	0.206	0.705	0.577	0.165	0.654	0.416
BASNet [23]	F	0.048	0.884	0.791	0.032	0.946	0.895	0.037	0.921	0.88	0.056	0.869	0.756
MINet [24]	F	0.037	0.917	0.828	0.029	0.96	0.909	0.033	0.953	0.924	0.056	0.873	0.755
KRN [13]	F	0.034	0.926	0.851	0.028	0.959	0.916	0.036	0.92	0.922	0.049	0.889	0.783
WSSA [4]	W	0.062	0.869	0.742	0.047	0.932	0.86	0.059	0.917	0.870	0.068	0.845	0.703
MFNet [17]	W	0.079	0.832	0.692	0.058	0.919	0.839	0.084	0.880	0.844	0.098	0.784	0.621
SCWS [25]	W	0.049	0.907	0.823	0.038	0.943	0.896	0.049	0.931	0.900	0.060	0.870	0.758
USPS [19]	U	0.068	0.85	0.747	0.045	0.923	0.88	0.067	0.893	0.873	0.062	0.848	0.738
UDASOD [20]	U	0.05	0.897	0.795	0.035	0.947	0.883	0.043	0.94	0.895	0.059	0.849	0.733
UMNet [21]	U	0.067	0.863	0.752	0.041	0.939	0.889	0.064	0.904	0.879	0.063	0.860	0.743
A2S [6]	U	0.069	0.847	0.729	0.041	0.936	0.868	0.056	0.921	0.882	0.079	0.818	0.688
A2SV2 [22]	U	**0.047**	0.903	0.81	0.037	0.948	0.903	**0.044**	**0.940**	**0.917**	**0.061**	**0.864**	0.746
OURS	U	0.048	**0.905**	**0.822**	**0.033**	**0.953**	**0.915**	0.048	0.936	0.916	0.064	0.862	**0.752**

4.4 Comparison Experiments

This section compares the method proposed in this paper with fully-supervised, weakly-supervised, and unsupervised methods for salient object detection, including: RBD [10], BASNet [23], MINet [24], KRN [13], USPS [19], UDASOD [20], A2S [6], A2SV2 [22], MFNet [17], SCWS [35], UMNet [21], USPS [19] and WSSA [4]. The effectiveness of each method is evaluated by comparing the saliency maps they generate, either using the original code or directly provided by the authors. The comparisons aim to highlight the performance gap between unsupervised methods, which do not require manual annotations, and other supervised approaches. Additionally, the section emphasizes the performance of the proposed method, which operates without any manual annotations. All methods are evaluated using the same evaluation code to ensure fairness.

Quantitative Analysis The assessments are shown in Table 1. "Method" indicates the model name. "Sup" denotes the supervision method of the model, where "T" represents traditional methods, "F" indicates fully-supervised methods, "W" stands for weakly-supervised methods, and "U" signifies unsupervised methods. Results in bold font represent the best performance among unsupervised methods.

Qualitative Analysis As shown in Figure 5, compared with the current mainstream weakly-supervised and unsupervised methods, the method proposed in this paper demonstrates significant advantages on various types of images. Particularly in the first to second rows of images, the method in this paper performs excellently in detecting the salient object "door", almost accurately completing the segmentation of the region while maintaining the complete edges and detailed features of the "door". Compared with previous methods, they have deficiencies in detecting the details and edges of the "door". Furthermore, the method in this paper can accurately segment salient objects in complex scenes, as shown in the third to fourth rows. Additionally, it can precisely segment salient objects when

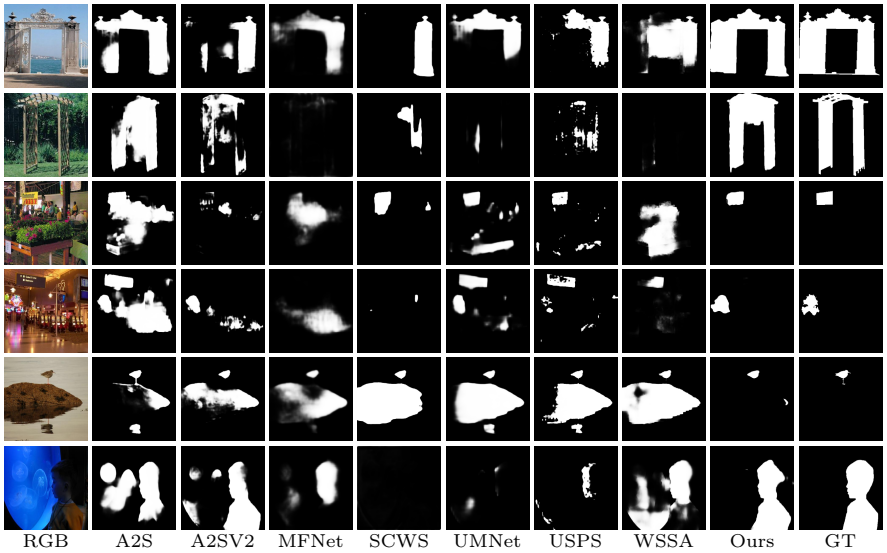

Fig. 5. Qualitative comparison of the methodology in this paper with other methods

they are small or when the input images have insufficient lighting. The above experimental results demonstrate the excellent performance of the method in this paper for salient object detection in complex tasks.

4.5 Ablation Studies

To evaluate the contributions of the various modules in the proposed method, this paper first established a baseline model. This model only uses CCAM and DCRF to generate pseudo-labels for supervision and excludes the Multi-feature Aggregation Module (MFA) and the Context Feature Interaction Module (CFI), serving as the baseline model. Subsequently, this paper incrementally added the proposed modules to the baseline model and analyzed the contributions of each module in detail. As shown in the results in Table 2, each module introduced into the model plays a decisive role in achieving the final excellent performance. It can be concluded that the method proposed in this paper makes significant contributions to salient object detection.

Table 2. Ablation experiments on DUT-OMRON dataset

MOCO	DINO	PSU	MFA	CFI	$F_\beta \uparrow$	$E_m \uparrow$
✓	×	×	×	×	0.716	0.835
×	✓	×	×	×	0.663	0.793
✓	✓	×	×	×	0.726	0.835
✓	×	✓	×	×	0.727	0.838
✓	✓	✓	×	×	0.731	0.840
✓	✓	✓	✓	×	0.743	0.848
✓	✓	✓	✓	✓	**0.752**	**0.862**

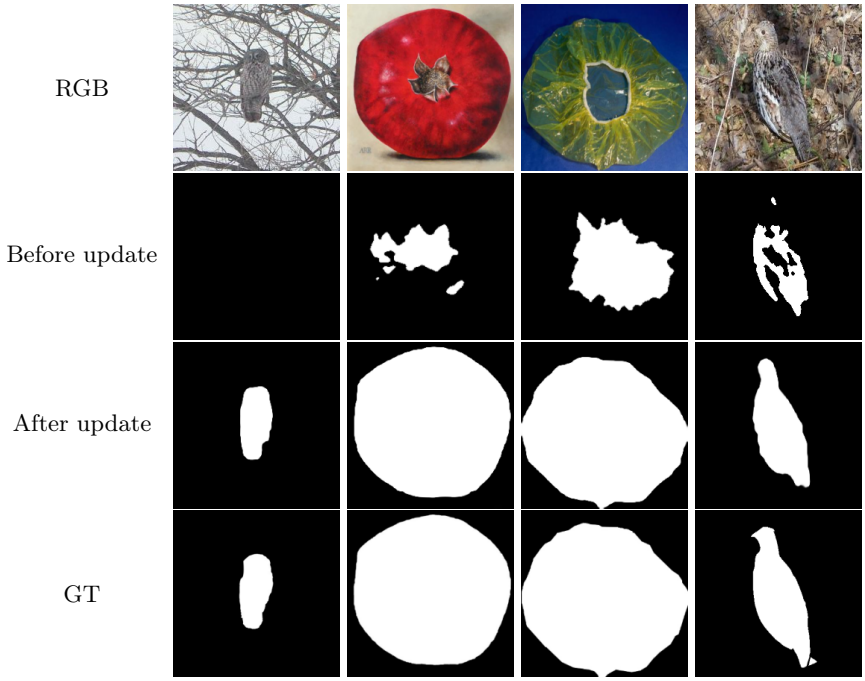

Fig. 6. Comparison of pseudo-labels before and after the update.

As shown in Figure 6, the visual differences between the pseudo-labels before and after updating are displayed. It is evident that the pseudo-labels updated using the self-updating method are closer to the ground-truth labels and better suited for the salient object detection task.

5 Conclusion

The comprehensive evaluation across multiple datasets demonstrates the robustness and effectiveness of the proposed method. Our approach consistently delivers competitive performance compared to both unsupervised and mainstream methods. Specifically, it matches the performance of fully-supervised and weakly-supervised methods on some datasets, while maintaining comparable results with mainstream methods on others. These findings highlight the potential of our method to bridge the gap between unsupervised and supervised learning in salient object detection. Future work will focus on optimizing the model architecture further and exploring its application in more diverse and complex scenarios.

References

1. Li, G., Xie, Y., Lin, L.: Weakly supervised salient object detection using image labels. In: AAAI, vol. 32, no. 1, pp. 7024–7031. AAAI Press, 2018.

2. Zhang, L., Zhang, J., Lin, Z., Lu, H., He, Y.: Capsal: Leveraging captioning to boost semantics for salient object detection. In: IEEE CVPR, pp. 6024–6033. IEEE, 2019.
3. Liu, Y., Wang, P., Cao, Y., Liang, Z., Lau, R.W.H.: Weakly-supervised salient object detection with saliency bounding boxes. IEEE TIP **30**, 4423–4435 (2021).
4. Zhang, J., Yu, X., Li, A., Song, P., Liu, B., Dai, Y.: Weakly-supervised salient object detection via scribble annotations. In: IEEE CVPR, pp. 12546–12555. IEEE, 2020.
5. Gao, S., Zhang, W., Wang, Y., Guo, Q., Zhang, C., He, Y., Zhang, W.: Weakly-supervised salient object detection using point supervision. In: AAAI, vol. 36, no. 1, pp. 670–678. AAAI Press, 2022.
6. Zhou, H., Chen, P., Yang, L., Xie, X., Lai, J.: Activation to saliency: Forming high-quality labels for unsupervised salient object detection. IEEE TCSVT **33**(2), 743–755 (2022).
7. Xie, J., Xiang, J., Chen, J., Hou, X., Zhao, X., Shen, L.: C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In: IEEE CVPR, pp. 989–998. IEEE, 2022.
8. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020).
9. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: IEEE International Conference on Computer Vision, pp. 9650–9660. IEEE, 2021.
10. Zhu, W., Liang, S., Wei, Y., Sun, J.: Saliency Optimization from Robust Background Detection. In: IEEE CVPR, pp. 2814–2821. IEEE, 2014.
11. Liu, J.-J., Hou, Q., Cheng, M.-M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: IEEE CVPR, pp. 3917–3926. IEEE, 2019.
12. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection. In: IEEE CVPR, pp. 9413–9422. IEEE, 2020.
13. Xu, B., Liang, H., Liang, R., Chen, P.: Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection. In: AAAI, vol. 35, no. 4, pp. 3004–3012. AAAI Press, 2021.
14. Liang, W., Ran, P., Bai, M., Liu, X., Githinji, P.B., Zhao, W., Qin, P.: External Prompt Features Enhanced Parameter-Efficient Fine-Tuning for Salient Object Detection. In: International Conference on Pattern Recognition, pp. 82–97. Springer, 2024.
15. Piao, Y., Wang, J., Zhang, M., Ma, Z., Lu, H.: To be Critical: Self-Calibrated Weakly Supervised Learning for Salient Object Detection. arXiv preprint arXiv:2109.01770 (2021).
16. Zhang, J., Yu, X., Li, A., Song, P., Liu, B., Dai, Y.: Weakly-supervised salient object detection via scribble annotations. In: IEEE CVPR, pp. 12546–12555. IEEE, 2020.
17. Piao, Y., Wang, J., Zhang, M., Lu, H.: Mfnet: Multi-filter directive network for weakly supervised salient object detection. In: IEEE International Conference on Computer Vision, pp. 4136–4145. IEEE, 2021.
18. Gao, S., Zhang, W., Wang, Y., Guo, Q., Zhang, C., He, Y., Zhang, W.: Weakly-Supervised Salient Object Detection Using Point Supervision. In: AAAI, 2022.
19. Nguyen, T., Dax, M., Mummadi, C.K., Ngo, N., Nguyen, T.H.P., Lou, Z., Brox, T.: Deepusps: Deep robust unsupervised saliency prediction via self-supervision. In: NIPS, vol. 32, 2019.

20. Yan, P., Wu, Z., Liu, M., Zeng, K., Lin, L., Li, G.: Unsupervised domain adaptive salient object detection through uncertainty-aware pseudo-label learning. In: AAAI, vol. 36, no. 3, pp. 3000–3008. AAAI Press, 2022.
21. Wang, Y., Zhang, W., Wang, L., Liu, T., Lu, H.: Multi-source uncertainty mining for deep unsupervised saliency detection. In: IEEE CVPR, pp. 11727–11736. IEEE, 2022.
22. Zhou, H., Qiao, B., Yang, L., Lai, J., Xie, X.: Texture-guided saliency distilling for unsupervised salient object detection. In: IEEE CVPR, pp. 7257–7267. IEEE, 2023.
23. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: IEEE CVPR, pp. 7479–7489. IEEE, 2019.
24. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection. In: IEEE CVPR, pp. 9413–9422. IEEE, 2020.
25. Yu, S., Zhang, B., Xiao, J., Lim, E.G.: Structure-consistent weakly supervised salient object detection with local saliency coherence. In: AAAI, vol. 35, no. 4, pp. 3234–3242. AAAI Press, 2021.
26. 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).
27. Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017).
28. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: IEEE CVPR, pp. 1597–1604. IEEE, 2009.
29. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: IEEE CVPR, pp. 733–740. IEEE, 2012.
30. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: IEEE CVPR, pp. 136–145. IEEE, 2017.
31. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: IEEE CVPR, pp. 1155–1162. IEEE, 2013.
32. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.-H.: Saliency detection via graph-based manifold ranking. In: IEEE CVPR, pp. 3166–3173. IEEE, 2013.
33. Li, G., Yu, Y.: Deep contrast learning for salient object detection. In: IEEE CVPR, pp. 478–487. IEEE, 2016.
34. Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
35. Yang, T., Wang, Y., Zhang, L., Qi, J., Lu, H.: Depth-inspired label mining for unsupervised rgb-d salient object detection. In: ACM Multimedia, pp. 5669–5677. ACM, 2022.