# CPDR: Towards Highly-Efficient Salient Object Detection via Crossed Post-decoder Refinement

Yijie Li[1]
yijieli2025@u.northwestern.edu
Hewei Wang[2]
heweiw@andrew.cmu.edu
Aggelos Katsaggelos[1]
a-katsaggelos@northwestern.edu

[1] Northwestern University
633 Clark St
Evanston, IL, USA

[2] Robotics Institute
Carnegie Mellon University
Pittsburgh, PA, USA

**Abstract**

Most of the current salient object detection approaches use deeper networks with large backbones to produce more accurate predictions, which results in a significant increase in computational complexity. A great number of network designs follow the pure UNet and Feature Pyramid Network (FPN) architecture which has limited feature extraction and aggregation ability which motivated us to design a lightweight post-decoder refinement module, the crossed post-decoder refinement (CPDR) to enhance the feature representation of a standard FPN or U-Net framework. Specifically, we introduce the Attention Down Sample Fusion (ADF), which employs channel attention mechanisms with attention maps generated by high-level representation to refine the low-level features, and Attention Up Sample Fusion (AUF), leveraging the low-level information to guide the high-level features through spatial attention. Additionally, we proposed the Dual Attention Cross Fusion (DACF) upon ADFs and AUFs, which reduces the number of parameters while maintaining the performance. Experiments on five benchmark datasets demonstrate that our method outperforms previous state-of-the-art approaches.

## 1 Introduction

Salient object detection (SOD) has rapidly evolved to become a cornerstone of modern computer vision, underpinning transformative advances across a diverse array of applications such as autonomous driving systems and robot exploration, where it enhances robotics real-time decision-making [7], facilitates object manipulation and human-robot interaction [6, 33]. Additionally, SOD plays a crucial role in professional-grade image editing by simplifying complex object isolation tasks [12]. The significance of SOD lies in its ability to prioritize regions in visual scenes that most attract human attention, thus serving as a foundational technology for higher-level computer vision tasks including scene understanding and adaptive compression. Contemporary approaches in SOD predominantly harness deep convolutional neural networks, drawing on sophisticated architectural innovations that have

shown remarkable success in feature representation and extraction. These methodologies often extend the foundational designs of U-Net [28] and Feature Pyramid Networks (FPN) [21], integrating multi-scale contextual information that significantly boosts the accuracy and robustness of saliency detection models.

However, despite these advances, there are several notable limitations in existing models:

- Deeper networks with large backbones and decoders significantly increase the computational burden, making them less feasible for deployment in resource-constrained environments.
- The traditional FPN and U-Net architectures, though widely used, exhibit limited capability in optimal feature aggregation and representation, potentially compromising the detection performance in complex scenes.

To ameliorate the aforementioned issues, we are motivated to propose a novel lightweight architecture, the Crossed Post-decoder Refinement (CPDR), which introduces a highly-efficient post-decoder refinement module to enhance the feature representation in standard FPN or U-Net frameworks. Our approach strategically employs channel attention mechanisms to refine the low-level feature maps, which are then intricately crossed with spatial attention maps derived from high-level features. This cross-attention scheme ensures a more effective and efficient feature integration and representation.

The primary contributions of the CPDR architecture are threefold:

- We introduce a novel lightweight post-decoder refinement technique, Attention Down-Sample Fusion (ADF) and Attention Up-Sample Fusion (AUF) that integrates channel and spatial attention mechanisms to significantly enhance the saliency detection capabilities of traditional networks without special-designed decoders.
- We propose a simplified CPDR module, Dual Attention Cross Fusion (DACF) with less computational complexity which is suitable for super-efficient salient object detection requirements, achieving 0.041 MAE with only 1.66M parameters.
- Our proposed CPDR model demonstrates superior performance on five benchmark datasets, surpassing previous state-of-the-art methods while maintaining lower computational complexity.

Through these innovations, CPDR not only improves the effectiveness of salient object detection but also addresses the practical challenges associated with deploying deep learning models in environments with stringent resource constraints.

## 2 Related Work

### Traditional Approaches in SOD

Prior to the rise of deep learning, traditional SOD methods were predominantly reliant on heuristic features and hand-crafted models. Classic techniques such as contrast-based methods exploited the distinct visual features of objects to distinguish them from their surroundings [10]. Region-based approaches took into account the homogeneity of regions to identify salient areas, often requiring significant feature engineering [3]. The most notable of these

early attempts was the frequency-tuned method which utilized the color and luminance features to detect salient regions in an image, marking the beginning of more sophisticated SOD methodologies [1]. Methods like UFO, which stands for Uniqueness, Focusness, and Objectness, offered early insights into the core attributes that make objects stand out [11]. A comprehensive benchmark by Borji et al. [2] provided a valuable evaluation of various traditional SOD algorithms, highlighting their strengths and limitations.

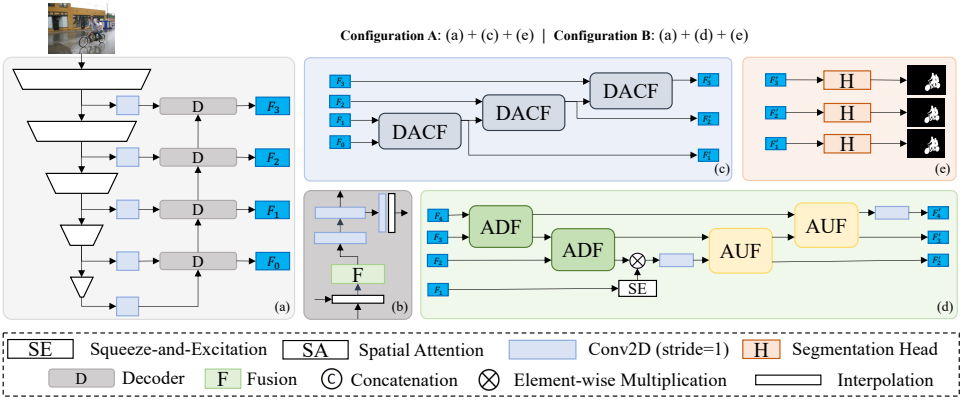


Figure 1: Overall Pipeline of CPDR

## Large-Scale Learning-based Models in SOD

In the realm of SOD, the emergence of deep learning has brought forth several large-scale models that stand out due to their intricate architectures and advanced feature extraction abilities. For instance, PoolNet leverages a multi-level feature aggregation strategy to refine saliency detection, offering impressive results across benchmark datasets [19]. Notably, recurrent strategies have been applied to SOD as well, as exemplified by Hu et al. [9]. Moreover, PiCANet proposed an innovative pixel-wise contextual attention mechanism [18], and Zhao et al. introduced the Pyramid Feature Attention Network (PFAN) [41], further enriching the landscape of SOD methodologies. BASNet introduces a boundary-aware inference mechanism that allows precise saliency mapping, often proving crucial for fine-grained object delineation [26]. Similarly, U2Net emphasizes a dual attention mechanism that separates salient features from complex backgrounds, demonstrating remarkable adaptability to varied scenes [39]. C2SNet employs a contour-to-saliency transferring method that predicts salient objects and their contours simultaneously, enhancing detection accuracy [31]. Continuing to advance the field, PurNet, a novel approach, adopts a purity-oriented structure to isolate salient objects against noisy backgrounds [42]. Lastly, Dual Attention Aggregating Network (DAANet) introduces a dual attention aggregating network, adding a new layer of feature refinement and aggregation for enhanced saliency detection [14].

## Lightweight Learning-based Models in SOD

With the increasing demand for real-time applications, there has been a focus on developing lightweight models that maintain a balance between efficiency and accuracy. The EDN model developed by Wu et al. in 2022, which stands for Extremely-Downsampled Network

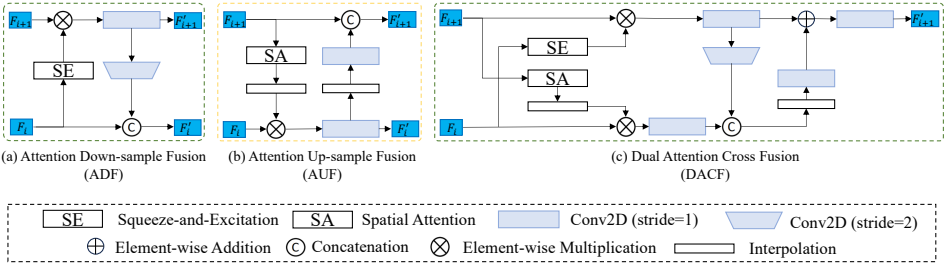


Figure 2: Demonstration of CPDR Modules

(EDN), showcases a high level of precision through novel feature compression techniques that do not compromise detection quality [35]. This model is particularly suitable for deployment in resource-constrained environments where computational efficiency is as vital as the accuracy of the saliency detection. In contrast, the Efficient Detection Network (EDN) model from earlier works like Mei et al. in 2017 refers to a different approach and should be clearly differentiated to avoid confusion [23]. The Cross-Attention Siamese Network (CASNet) focuses on video SOD by efficiently processing temporal information across frames [22]. Additionally, the Hierarchical Visual Perception Network (HVPNet) [20] and Stereoscopically Attentive Multi-scale Network (SAMNet) [21] enhance lightweight saliency detection by utilizing hierarchical and attentive multi-scale processing. The CII18 model, which rethinks traditional U-shape structures, further improves performance and efficiency in saliency detection [17]. These models, including HVPNet, SAMNet, CII18, and the two distinct EDN models, exemplify the trend toward solutions that are both computationally efficient and highly accurate in detecting salient objects.

# 3 Crossed Post-decoder Refinement (CPDR)

## 3.1 Overview of CPDR

As shown in Figure 1, our proposed Crossed Post-decoder Refinement (CPDR) serves as a refinement module after the decoders. In Figure 1 (a), we use standard UNet [28] and Feature Pyramid Network (FPN) [16] with modified MobileNetV2 [29] and EfficientNet [30] B0/B3 backbones for experiments. The structure of the decoders is shown in Figure 1 (b) which only contains three convolution layers without special design. In our smallest version (Ours-S), we modified the original MobileNetV2 [29] by removing the last two blocks which gives us a simplified backbone with only 1.4 million parameters. In this configuration, we use FPN as the encoder-decoder network, with a Dual Attention Cross Fusion (DACF) as the refinement module, shown in Figure 1 (c). As for our medium (Ours-M) and large version (Ours-L), we use EfficientNet-B0 and EfficientNet-B3 as backbone, UNet as encoder-decoder network, and their refinement module is a U-shaped module, illustrated in Figure 1 (d).

## 3.2 Attention Down-Sample Fusion (ADF)

The channels attention mechanism is widely used in computer vision tasks, which emphasize certain channels over others. However, most of the current channels attention is applied to

the same feature maps, focusing on single-level feature re-weight. Motivated by Squeeze-and-Excitation (SE) [8], we proposed the Attention Down-Sample Fusion (ADF) with cross-level channels attention for more effective feature fusion, shown in Figure 2 (a). The basic idea of ADF is the channel dimension of a high-level feature consists of richer information, which can be utilized to create an attention map for lower-level features. Considering a high-level feature $F_i$ and a low-level feature $F_{i+1}$, we compute the $F^{'}_{i+1}$ by applying the channels attention followed by a convolution, while $F^{'}_{i}$ can be calculated based on $F^{'}_{i+1}$ output:

$$F^{'}_{i+1} = \text{Conv2D}(\text{SE}(F_i) \odot F_{i+1}) \tag{1}$$

$$F^{'}_{i} = \text{Concat}(\{\text{F}_\text{i}, \text{Down}(\text{F}^{'}_{\text{i+1}})\}) \tag{2}$$

where Down is a convolution layer with a stride equal to 2. The ADF provided a simple way to effectively fuse the low-level feature with high-level features under a 'cross-attention' manner.

### 3.3 Attention Up-Sample Fusion (AUF)

However, as shown in Figure 1 (d), we only consider $F^{'}_{3}$ as the final output, which means features fused by ADFs don't have any contribution to the final output, so we proposed the Attention Up-Sample Fusion (AUF) to match these ADFs, shown in Figure 2 (b). Sharing a similar idea of ADFs, AUF considers that lower-level feature with higher resolution contains more accurate spatial information which can be used to create attention for high-level features. Then, the output $F^{'}_{i}$ can be calculated by:

$$\text{Conv2D}(\text{Interpolate}(\text{SA}(F_{i+1})) \odot F_i) \tag{3}$$

where SA is the CBAM [34] spatial attention and Interpolate represent the bilinear interpolation. The output $F^{'}_{i+1}$ can be computed through:

$$\text{Concat}(\{F_{i+1}, \text{Conv2D}(\text{Interpolate}(F^{'}_{i})\})) \tag{4}$$

The AUFs successfully fused all feature maps generated by ADFs to the final output by utilizing the guidance of spatial attention.

### 3.4 Dual Attention Cross Fusion (DACF)

In order to further reduce the computational complexity, we introduce Dual Attention Cross Fusion (DACF), which combines an ADF and an AUF with some of the $3 \times 3$ convolution layers replaced by $1 \times 1$ convolution layers, as shown in Figure 2 (c). The DACF separately applies channel attention and spatial attention to low-level feature maps and high-level future maps, which is different from the ADF and AUF pipeline. The weighted $F_{i+1}$ ($wF_{i+1}$) and $F_i$ ($wF_i$) can be directly calculated by the same method used in ADF and AUF. As for the fusion, we first concatenate the two weighted feature maps though $\text{Concat}(\{\text{Down}(wF_{i+1}), wF_i\})$ to produce $C_i$. Finally, we aggregate $C_i$ to generate output $F^{'}_{i+1}$ by computing:

$$\text{Conv2D}(\text{Conv2D}(\text{interpolate}(C_i)) + wF_{i+1}) \tag{5}$$

The DACF simplified the ADF+AUF pipeline with less computation complexity which makes DACF suitable for the small configuration, Ours-S.

### 3.5 Objective Functions

We use DICE loss and Intersection over Union (IoU) loss for training. The DICE loss can be formulated as:

$$\mathcal{L}_{\text{DICE}} = 1 - \frac{\sum_{m,n} p \odot y + \varepsilon}{\sum_{m,n} p + \sum_{m,n} y + \varepsilon} \tag{6}$$

and IOU loss is:

$$\mathcal{L}_{\text{IOU}} = 1 - \frac{\sum_{m,n} p \odot y + \varepsilon}{\sum_{m,n} p + \sum_{m,n} y - \sum_{m,n} p \odot y + \varepsilon} \tag{7}$$

where $p$ represent the prediction and $y$ represent the ground truth. $\varepsilon$ indicate the smoothing coefficient and $\odot$ is the element-wise multiplication. Then, the total loss functions can be formulated by:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DICE}} + \mathcal{L}_{\text{IOU}} \tag{8}$$

In order to speed up the convergence speed, we use the deep-supervision strategy which means the $\mathcal{L}_{\text{total}}$ will be applied to the three predictions from three stages and ground truth will be interpolated to match the size of each prediction.

## 4 Experiments

### 4.1 Datasets

We utilize the DUTS-TR dataset [32] to train our DAANet, assessing its performance on multiple benchmarks including DUTS-TE [32], HKU-IS [13], ECSSD [36], PASCAL-S [15], and DUT-OMRON [37]. The training dataset, DUTS-TR, comprises 10,553 samples. For evaluation, DUTS-TE provides 5,019 images, which is the most widely-used test set for SOD task. HKU-IS features 4,447 image pairs, and ECSSD offers 1,000 test images. Additionally, PASCAL-S contains 850 images and DUT-OMRON includes 5,168 images for testing.

### 4.2 Evaluation Metrics

We employ five metrics to assess the performance of CPDR:

- Mean Absolute Error (MAE) [25], defined as $\frac{1}{N \times M} \sum_{n=1}^{N} \sum_{m=1}^{M} |t_m - p_m|$, which quantifies the average per-pixel discrepancy between the prediction and ground truth.
- Mean F-measure ($F_\beta^m$), serves as the weighted harmonic mean between precision and recall, which is formulated by $\frac{(1+\beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$, where $\beta^2 = 0.3$.
- Mean E-measure ($E_\phi^m$) [5], combines the single pixel values with the global-level mean value, which can be calculated as $\frac{1}{N \times M} \sum_{n=1}^{N} \sum_{m=1}^{M} \theta(\phi)$, where $\phi$ is the alignment matrix and $\theta(\phi)$ is the enhanced alignment matrix.

- $S$-measure ($S_m$) [4], designed to quantitatively evaluate structural similarity in a manner that closely aligns with human visual perception. This metric is meticulously computed using the formula $S_m = m \cdot s_o + (1-m) \cdot s_r$, where $s_o$ and $s_r$ denote the object-aware and region-aware structural similarities, respectively. The parameter $m$ is strategically set to 0.5 to balance the influence of both components.

- Weighted F-measure ($F_\beta^\omega$) [5] offers a nuanced adaptation of the traditional $F_\beta$ measure by emphasizing precision and recall differently based on the spatial characteristics of the errors. It is calculated as:

$$F_\beta^\omega = \frac{(1+\beta^2) \cdot (\omega_p \cdot \text{Precision}) \cdot (\omega_r \cdot \text{Recall})}{\beta^2 \cdot (\omega_p \cdot \text{Precision}) + (\omega_r \cdot \text{Recall})},$$

  where $\omega_p$ and $\omega_r$ are weighting functions that adaptively modify the influence of precision and recall, respectively, based on localized error characteristics and $\beta^2 = 0.3$. This metric considers the specific locations and neighborhood context, making it especially suitable for non-binary evaluation scenarios.

## 4.3 Implementation Detail

We perform the training on a single NVIDIA Tesla V100-SXM2 16GB GPU. We train our models on DUTS-TR [32] dataset. The training batch size for all experiments is set to 16 and 40 epochs for the model with MobileNetV2 backbone and 20 epochs with EfficientNet backbones. We use Adam as the training optimizer. A poly learning rate scheduler with a linear warm-up is adopted, we warm up the training with 5 epochs, and $\gamma$ for poly learning rate decay is set to 3 for MobileNetV2 backbone and 5 for EfficientNet backbones. Images are resized to $(256, 256)$ and only a random horizontal flip is adopted for data augmentation.

## 4.4 Ablation Study

Table 1: Ablation study on module compositions, loss functions, and backbone networks

| No. | Backbone | Arch | CPDR Config | | | Loss Function | | | #Params | DUTS-TE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | ADF | AUF | DACF | DICE | BCE | IOU | (M) | $\mathcal{M}\downarrow$ | $F_\beta^m\uparrow$ | $F_\beta^\omega\uparrow$ | $S_m\uparrow$ | $E_\phi^m\uparrow$ |
| 1 | MobileNetV2* | FPN | | | | | ✓ | ✓ | 1.58 | .048 | .800 | .776 | .850 | .899 |
| 2 | | FPN | ✓ | ✓ | | | ✓ | ✓ | 1.72 | .044 | .812 | .790 | .856 | .903 |
| 3 | | UNet | ✓ | ✓ | | | ✓ | ✓ | 1.82 | .044 | .814 | .793 | .857 | .905 |
| 4 | | UNet | | | ✓ | | ✓ | ✓ | 1.75 | .044 | .813 | .791 | .855 | .904 |
| 5 | | FPN | | | ✓ | | ✓ | ✓ | 1.66 | .044 | .810 | .787 | .854 | .902 |
| 6 | | FPN | | | ✓ | ✓ | | ✓ | 1.66 | .041 | .822 | .803 | .851 | .910 |
| 7 | EfficientNet-B0 | FPN | ✓ | ✓ | | | ✓ | ✓ | 4.54 | .041 | .826 | .808 | .867 | .912 |
| 8 | | UNet | ✓ | ✓ | | | ✓ | ✓ | 4.64 | .040 | .828 | .811 | .869 | .912 |
| 9 | | UNet | ✓ | ✓ | | ✓ | | ✓ | 4.64 | .038 | .839 | .825 | .864 | .922 |
| 10 | EffcientNet-B3 | UNet | ✓ | ✓ | | ✓ | | ✓ | 11.36 | **.034** | **.853** | **.842** | **.875** | **.931** |

To demonstrate the effectiveness of our Crossed Post-decoder Refinement (CPDR) module, we conduct a comprehensive ablation study on the backbone networks, encoder-decoder architectures, module compositions, and loss functions, as shown in Table 1 on DUTS-TE [32] dataset with five metrics: MAE ($\mathcal{M}$), mean F-measure ($F_\beta^m$), weighted F-measure ($F_\beta^\omega$), S-measure ($S_m$), and mean E-measure ($E_\phi^m$). We consider an FPN architecture with our modified MobileNetV2 [29] backbone trained with BCE and IOU loss as the baseline (No. 1), which has 1.58M parameters. It can achieve an MAE of 0.048 and a weighted F-measure of 0.8. In experiment No. 2, we maintain the FPN structure and equip the pipeline with

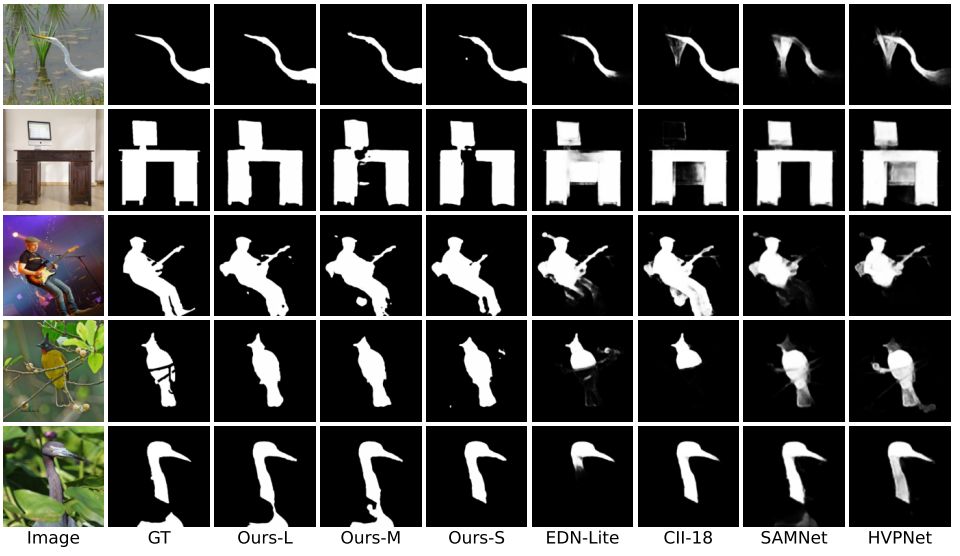


Figure 3: Qualitative comparison on DUTS-TE dataset between our methods and previous lightweight state-of-the-art approaches

ADFs and AUFs, with only a 0.14M increase in parameters amount, but the performance on all five metrics is significantly improved with a new MAE of 0.044. By comparing No. 2 and No. 3, we find that adopting UNet [28] instead of FPN [41] will increase the number of parameters but the performance improvement is not significant. In experiments No. 4 and No. 5, we compare the performance of UNet and FPN structure with DACF, and we observe that the parameters amount is reduced without significant performance drop. In No. 6, we adopt DICE+IOU as loss functions. Considering many previous approaches have demonstrated the effectiveness of IOU loss, we only compare the results of No. 6 with No. 5 (BCE+IOU) which shows that DICE+IOU is better than BCE+IOU which gets an MAE of 0.041 with only 1.66M parameters using FPN and DACF. From No. 7 to No. 9, we conduct experiments using EfficientNet-B0 backbone [30], the results illustrate that UNet structure with ADF and AUF trained with DICE and IOU loss can achieve the best performance. Finally, in No. 10, utilizing the EfficientNet-B3 backbone, our approach gets an MAE of 0.034 with only 11M parameters.

## 4.5 Qualitative Evaluation

Figure 3 illustrates the qualitative results on DUTS-TE dataset. We can observe that our model (Ours-L, Ours-M, and Ours-S), achieves superior segmentation performance and completeness. Our models adeptly handle diverse and challenging imaging conditions. They excel in scenarios involving complex structures in row 2 and 4, multiple objects such as row 3, where other models often lose detail or misinterpret the boundaries, and cluttered backgrounds like row 5, as well as in distinguishing small and intricate details such as those seen in rows 4 and 5. These results demonstrate the capability of our CPDR models to deliver high-quality segmentation outputs consistently across various challenging scenarios.

## 4.6 Quantitative Evaluation

Table 2 and F-measure curves shown in Figure 4 provide a comprehensive quantitative comparison and performance visualization of various SOD models across multiple datasets. Notably, three versions of a new model, designated as Ours (S), Ours (M), and Ours (L), demonstrate varying levels of efficiency and effectiveness. Ours (S) achieves significant compactness with only 1.66M parameters and 1.02 MACs, yet maintains competitive performance metrics. Ours (M) strikes an optimal balance between speed and accuracy with its 4.64M parameters, outperforming many existing models. Ours (L) excels in performance, utilizing 11.36M parameters to surpass nearly all advanced SOD methods, thus establishing new benchmarks in both efficiency and effectiveness.

Table 2: Quantitative comparison with state-of-the-art approaches

| Methods | #Params | MACs | ECSSD | | | PASCAL-S | | | DUTS-TE | | | HKU-IS | | | DUT-OMRON | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (M) | (G) | $\mathcal{M}\downarrow$ | $F_\beta^m\uparrow$ | $E_\phi^m\uparrow$ | $\mathcal{M}\downarrow$ | $F_\beta^m\uparrow$ | $E_\phi^m\uparrow$ | $\mathcal{M}\downarrow$ | $F_\beta^m\uparrow$ | $E_\phi^m\uparrow$ | $\mathcal{M}\downarrow$ | $F_\beta^m\uparrow$ | $E_\phi^m\uparrow$ | $\mathcal{M}\downarrow$ | $F_\beta^m\uparrow$ | $E_\phi^m\uparrow$ |
| Conventional CNN Models | | | | | | | | | | | | | | | | | |
| BASNet$_{19}$ [cvpr] | 87.06 | 127.36 | .037 | .917 | .943 | .076 | .818 | .879 | .048 | .823 | .896 | .032 | .902 | .943 | .056 | .767 | .865 |
| MINet$_{20}$ [cvpr] | 126.38 | 87.11 | .033 | .923 | .950 | .064 | .830 | .896 | .037 | .844 | .917 | .029 | .909 | .952 | .056 | .757 | .860 |
| GateNet$_{20}$ [eccv] | 128.63 | 162.13 | .033 | .927 | .948 | .062 | .844 | .901 | .037 | .851 | .917 | .029 | .916 | .952 | .054 | .770 | .865 |
| CII50$_{21}$ [cvpr] | 24.48 | 11.60 | .033 | .927 | .948 | .062 | .844 | .901 | .037 | .851 | .917 | .029 | .916 | .952 | **.054** | .770 | .865 |
| EDN$_{22}$ [tip] | 42.85 | 20.45 | .032 | **.930** | .951 | .062 | **.849** | .902 | .035 | .863 | .925 | .026 | .920 | .955 | .049 | **.788** | **.877** |
| ICON$_{22}$ [icml] | 30.09 | 20.91 | .032 | .928 | .954 | .064 | .838 | .899 | .037 | .853 | .924 | .029 | .912 | .953 | .057 | .779 | .876 |
| M$^3$Net$_{23}$ [cvpr] | 34.61 | 18.83 | **.029** | .926 | **.955** | **.060** | .844 | **.904** | **.037** | **.863** | **.927** | **.026** | **.920** | **.959** | .061 | .784 | .871 |
| Lightweight CNN Models | | | | | | | | | | | | | | | | | |
| HVPNet$_{21}$ [tip] | **1.24** | 1.1 | .052 | .882 | .911 | .089 | .783 | .844 | .058 | .772 | .859 | .044 | .867 | .913 | .065 | .736 | .839 |
| SAMNet$_{21}$ [cvpr] | 1.33 | **0.5** | .050 | .883 | .916 | .092 | .777 | .838 | .058 | .768 | .859 | .045 | .864 | .911 | .065 | .734 | .840 |
| CII18$_{21}$ [cvpr] | 11.89 | 8.48 | .039 | .913 | .939 | .068 | .824 | .888 | .043 | .831 | .904 | .032 | .904 | .945 | .058 | .747 | .849 |
| EDN(Lite)$_{22}$ [tip] | 1.80 | 1.14 | .042 | .910 | .933 | .073 | .818 | .877 | .045 | .819 | .895 | .034 | .897 | .937 | .057 | .746 | .848 |
| Ours(S) | 1.66 | 1.02 | .044 | .901 | .932 | .067 | .820 | .888 | .041 | .822 | .910 | .032 | .897 | .945 | .055 | .750 | .862 |
| Ours(M) | 4.64 | 2.67 | .037 | 914 | .944 | .064 | .830 | .898 | .038 | .839 | .922 | .030 | .903 | .951 | .052 | .764 | .871 |
| Ours(L) | 11.36 | 3.25 | **.033** | **.921** | **.951** | **.061** | **.836** | **.905** | **.034** | **.853** | **.931** | **.028** | **.908** | **.954** | **.048** | **.782** | **.883** |

The F-measure curves highlight that Ours (L) consistently maintains high scores across all thresholds, underlining its robust performance, while other models like HVPNet and EDN-Lite show a steeper decline in F-measure as the threshold increases, indicating reduced performance at higher strictness levels. This analysis not only benchmarks the capabilities of these models but also underscores the advancements in computational efficiency and algorithmic precision in the field of visual perception.

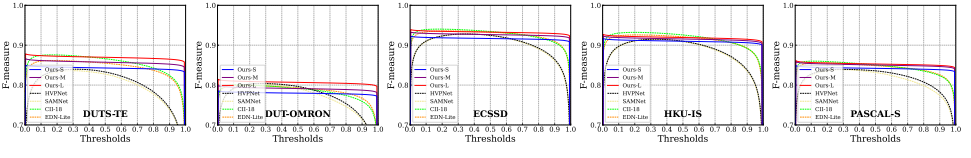


Figure 4: Illustration of F-measure curves on five benchmark datasets

# 5 Conclusions and Future Work

In this paper, we introduced the Crossed Post-decoder Refinement (CPDR) architecture, a new approach to salient object detection that significantly enhances the performance of traditional models like U-Net and Feature Pyramid Networks. By incorporating a lightweight post-decoder refinement module that utilizes cross-level feature aggregation with both channel and spatial attention mechanisms, our model achieves superior saliency detection capabilities while maintaining reduced computational complexity. The CPDR model not only

excels in benchmark tests against state-of-the-art methods but also offers a practical solution for deployment in environments with limited computational resources. For future work, an essential focus will be on optimizing the CPDR architecture for handling high-resolution images effectively while maintaining both accuracy and efficiency, since the increase of image resolution at the user-end results from the rapid development of imaging technology. This involves exploring strategies such as advanced down-sampling techniques to manage the increased computational load without losing significant image details crucial for accurate detection. Additionally, we are planning to further validate the universality of our approach, aiming to bring researchers a new plug-to-play module that is suitable for not only salient object detection but other widely studied computer vision tasks as well.

## Acknowlegdgement

The authors acknowledge the AI in Multimedia – Image and Video Processing Laboratory (AIM-IVPL) for computing support and advice.

## References

[1] R. Achanta et al. Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[2] A. Borji, M. M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 414–423, 2015.

[3] M. Cheng et al. Global contrast based salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[4] 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 (ICCV)*, Oct 2017.

[5] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *Scientia Sinica Informationis*, 6 (6):5, 2021.

[6] Ricardo Garcia, Georges Toulminet, and José M Armingol. A survey of image processing techniques for robotics. *RAM. Revista de Acústica*, 46(1):1–9, 2015.

[7] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3): 362–386, 2020.

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[9] Y. Hu et al. Recurrently aggregating deep features for salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[10] L. Itti et al. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998.

[11] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li. Salient region detection by ufo: Uniqueness, focusness and objectness. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1976–1983, 2013.

[12] Manish Kumar, Vijaypal Singh Dhaka, and Partha Pratim Roy. A survey on salient object detection using deep learning. *Computer Science Review*, 35:100202, 2019.

[13] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5455–5463, 2015.

[14] Yijie Li, Hewei Wang, Zhenqi Li, Shaofan Wang, Soumyabrata Dev, and Guoyu Zuo. DAANet: Dual Attention Aggregating Network for Salient Object Detection. In *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1–7. IEEE, 2023.

[15] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 280–287, 2014.

[16] 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 (CVPR)*, pages 2117–2125, 2017.

[17] Jiang-Jiang Liu, Zhi-Ang Liu, Pai Peng, and Ming-Ming Cheng. Rethinking the u-shape structure for salient object detection. *IEEE Transactions on Image Processing*, 30:9030–9042, 2021.

[18] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3089–3098, 2018.

[19] X. Liu et al. Poolnet: A simple but effective baseline for salient object detection. In *European Conference on Computer Vision (ECCV)*, volume 123, 2019.

[20] Yun Liu, Yu-Chao Gu, Xin-Yu Zhang, Weiwei Wang, and Ming-Ming Cheng. Lightweight salient object detection via hierarchical visual perception learning. *IEEE transactions on cybernetics*, 51(9):4439–4449, 2020.

[21] Yun Liu, Xin-Yu Zhang, Jia-Wang Bian, Le Zhang, and Ming-Ming Cheng. Samnet: Stereoscopically attentive multi-scale network for lightweight salient object detection. *IEEE Transactions on Image Processing*, 30:3804–3814, 2021.

[22] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Z. Li, and P. M. Jodoin. Casnet: A cross-attention siamese network for video salient object detection. *IEEE Transactions on Image Processing*, 29:4922–4933, 2020.

[23] K. Mei et al. Edn: A deep network for efficient salient object detection. *IEEE Transactions on Cybernetics*, 2017.

[24] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9413–9422, 2020.

[25] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 733–740, 2012.

[26] X. Qin et al. Basnet: Boundary-aware salient object detection. *IEEE Transactions on Image Processing*, 29, 2020.

[27] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7479–7489, 2019.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[29] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[30] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[31] L. Wang et al. C2snet: A deep network for sequential salient object detection. *IEEE Transactions on Image Processing*, 2019.

[32] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 136–145, 2017.

[33] Zihan Wang, Bowen Li, Chen Wang, and Sebastian Scherer. AirShot: Efficient few-shot detection for autonomous exploration. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024. URL https://arxiv.org/pdf/2404.05069.pdf.

[34] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[35] Yu-Huan Wu, Yun Liu, Le Zhang, Ming-Ming Cheng, and Bo Ren. Edn: Salient object detection via extremely-downsampled network. *IEEE Transactions on Image Processing*, 31:3125–3136, 2022.

[36] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1162, 2013.

[37] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3166–3173, 2013.

[38] Yao Yuan, Pan Gao, and XiaoYang Tan. $M^3$Net: Multilevel, mixed and multistage attention network for salient object detection. *arXiv preprint arXiv:2309.08365*, 2023.

[39] Z. Zhang et al. U2net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 2021.

[40] J. Zhao et al. Suppress and balance: A simple gated network for salient object detection. *ECCV*, 2020.

[41] T. Zhao et al. Pyramid feature attention network for saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[42] P. Zhou et al. Purnet: Pure salient object detection with purified structure. In *International Conference on Computer Vision (ICCV)*, 2020.

[43] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3738–3752, 2022.