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

Yijie Li[1], **Hewei Wang**[2], **Aggelos Katsaggelos**[1]

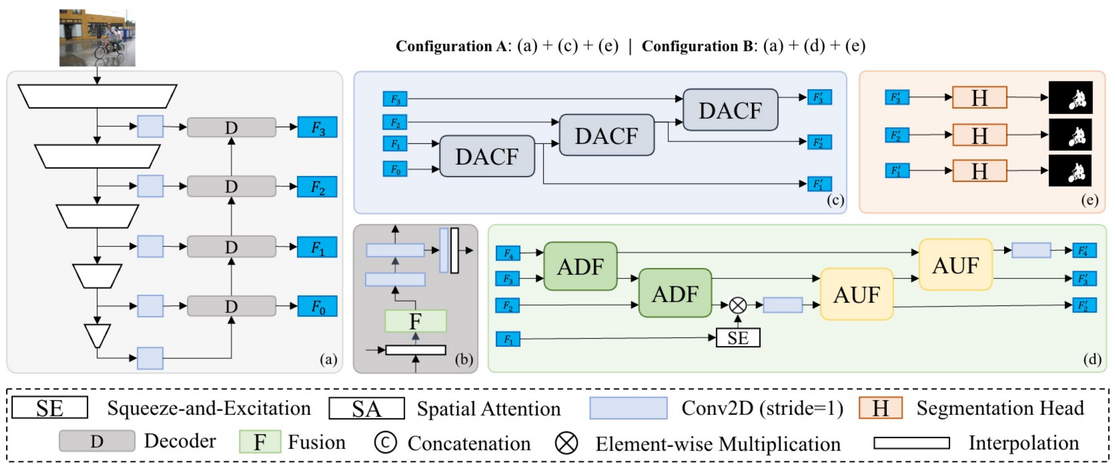
[1]Northwestern University, [2]Carnegie Mellon University

BMVC 2024

Overview

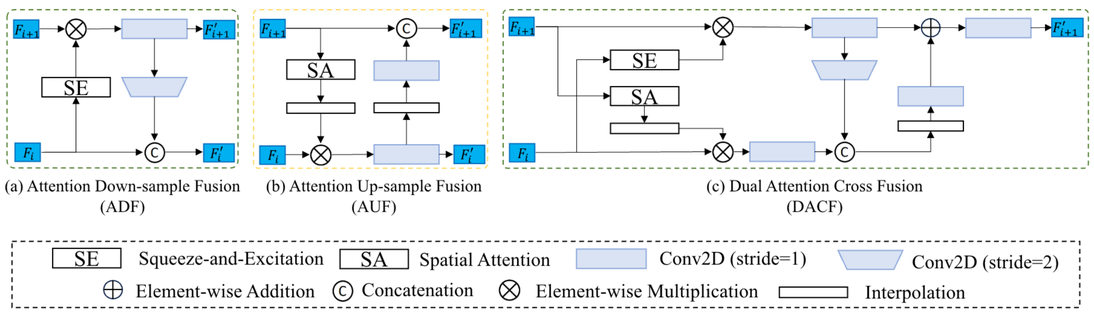
Salient object detection (SOD) identifies visually distinct objects in an image, which is critical for tasks like segmentation and scene analysis. Most SOD methods use deep networks with heavy backbones, increasing computational cost. Common architectures like U-Net and FPN have limited feature extraction and aggregation capabilities. To address this, we propose a lightweight Crossed Post-Decoder Refinement (CPDR) module, which enhances feature representation in FPN or U-Net frameworks, improving detection accuracy without added complexity.

Overall Pipeline



Our proposed Crossed Post-decoder Refinement (CPDR) serves as a refinement module after an encoder-decoder architecture. The Attention Down Sample Fusion (ADF) and the Attention Up Sample Fusion (AUF) are the main components of our medium and large configuration, while the Dual Attention Cross Fusion (DACF) is designed for the lightest version.

CPDR Modules



a. Attention Down Sample Fusion (ADF)

Motivated by the Squeeze-and-Excitation (SE) module, we proposed the Attention Down-Sample Fusion (ADF) with cross-level channels attention for more effective feature fusion. 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, it can be formulated as:

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

$$F'_i = \mathrm{Concat}(\{\mathrm{F_i}, \mathrm{Down}(\mathrm{F'_{i+1}})\})$$

b. Attention Up Sample Fusion (AUF)

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. This procedure can be formulated as:

$$F'_i = \mathrm{Conv2D}(\mathrm{Interpolate}(\mathrm{SA}(F_{i+1})) \odot F_i)$$

$$F'_{i+1} = \mathrm{Concat}(\{F_{i+1}, \mathrm{Conv2D}(\mathrm{Interpolate}(F'_i))\})$$

where SA is the Spatial Attention module.

c. 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 3x3 convolution layers replaced by 1x1 convolution layers. The DACF separately applies channel attention and spatial attention to low-level feature maps and high-level feature maps, which is different from the ADF and AUF pipeline.

Objective Functions

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

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

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{DICE}} + \mathcal{L}_{\mathrm{IOU}}$$

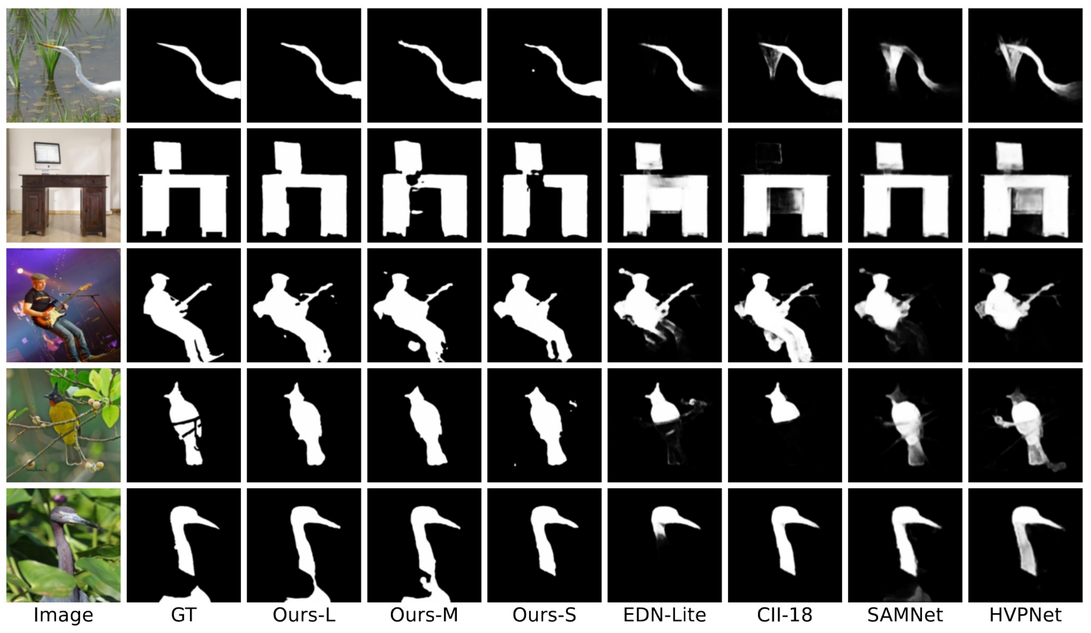
Ablation Study

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 on DUTS-TE dataset with five metrics: MAE, mean F-measure, weighted F-measure, S-measure, and mean E-measure.

Qualitative Results



Quantitative Results

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} [27]	87.06	127.36	.037	.917	.943	.076	.818	.879	.048	.823	.896	.032	.902	.943	.056	.767	.865
MINet_{20} [24]	126.38	87.11	.033	.923	.950	.064	.830	.896	.037	.844	.917	.029	.909	.952	.056	.757	.860
$\mathrm{GateNet}_{20}$ [40]	128.63	162.13	.033	.927	.948	.062	.844	.901	.037	.851	.917	.029	.916	.952	.054	.770	.865
$\mathrm{CII50}_{21}$ [17]	24.48	11.60	.033	.927	.948	.062	.844	.901	.037	.851	.917	.029	.916	.952	**.054**	.770	.865
EDN_{22} [35]	42.85	20.45	.032	**.930**	.951	.062	**.849**	.902	.035	.863	.925	.026	.920	.955	.049	**.788**	**.877**
ICON_{22} [43]	30.09	20.91	.032	.928	.954	.064	.838	.899	.037	.853	.924	.029	.912	.953	.057	.779	.876
$\mathrm{M^3Net}_{23}$ [38]	34.61	18.83	**.029**	.926	**.955**	**.060**	.844	**.904**	**.037**	**.863**	**.927**	**.026**	**.920**	**.959**	.061	.784	.871
Lightweight CNN Models																	
HVPNet_{21} [20]	**1.24**	1.1	.052	.882	.911	.089	.783	.844	.058	.772	.859	.044	.867	.913	.065	.736	.839
SAMNet_{21} [21]	1.33	**0.5**	.050	.883	.916	.092	.777	.838	.058	.768	.859	.045	.864	.911	.065	.734	.840
$\mathrm{CII18}_{21}$ [17]	11.89	8.48	.039	.913	.939	.068	.824	.888	.043	.831	.904	.032	.904	.945	.058	.747	.849
$\mathrm{EDN(Lite)}_{22}$ [35]	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**

We conducted a comprehensive quantitative comparison between our approach and recent years' state-of-the-art approaches on the five widely used salient object detection benchmark datasets, which demonstrates the advancement of our method.