LEARNING TO COMPLEMENT: RETHINKING FOCAL SELECTION FOR LIGHT FIELD SALIENT OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Light field salient object detection benefits from complementary focal cues, but independent slice scoring can favor redundant observations and overlook slices that become informative when combined. Effective fusion further requires accounting for dependencies between the selected focal evidence and all-focus features. We propose a network that integrates complementarity-aware focal selection with relation-conditioned state-space fusion. Specifically, Complementarity-Aware Focal Selection and Aggregation (CFSA) estimates the conditional prediction gains of individual slices and slice pairs given the accumulated evidence. Its pairwise look-ahead captures complementary combinations whose predictive value is underestimated by individual scoring. The estimated gains guide progressive evidence selection and aggregation, with adaptive stopping under an encoding budget. Relation-Conditioned State-Space Fusion (RCSF) then models all-focus and focal features as observations of a shared spatial state. Cross-modal relations determine the observation mappings and weights in a regularized joint correction. The resulting closed-form update jointly incorporates the observation residuals while accounting for overlap between observation directions, allowing cross-modal dependencies to shape state correction and subsequent spatial propagation. Extensive experiments on four public light field salient object detection benchmarks demonstrate that the proposed method outperforms state-of-the-art approaches.

1 INTRODUCTION

The value of a focal observation depends on what it adds to the available evidence. In light field salient object detection (LFSOD), observations at different focal planes provide cues for separating salient objects from visually similar surroundings Li et al. (2014), while an all-focus image supplies the overall scene appearance. Figure 1(a) illustrates focal slices and their corresponding saliency responses. Individually informative slices may repeatedly describe structures already represented, whereas observations that are ambiguous on their own may jointly clarify a foreground–background relationship. Effective focal selection therefore requires understanding how observations complement one another as evidence accumulates.

Existing LFSOD methods have made substantial progress through guided stack refinement Yuan et al. (2023), Transformer-based interaction Jiang et al. (2024), and state-space modeling of inter-slice and cross-modal dependencies Liu et al. (2024); Geng et al. (2025). Explicit selection has also been explored through foreground-similarity retrieval Lv et al. (2023) and variance-maximized key-slice selection Han et al. (2025). These advances establish the importance of informative observations and their relationships. However, relevance and diversity do not directly determine an observation's contribution to the current saliency prediction. A relevant slice may repeat evidence already represented, while a diverse slice may differ mainly in regions unrelated to the salient object. We make this conditional predictive utility an explicit criterion for focal selection.

Figure 1(b,c) contrasts dense aggregation with slice-wise selection. Beyond this distinction, conditional predictive utility exposes two challenges. Redundancy requires a selector to reassess candidate slices as evidence accumulates: an initially useful observation may contribute little after similar ev-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

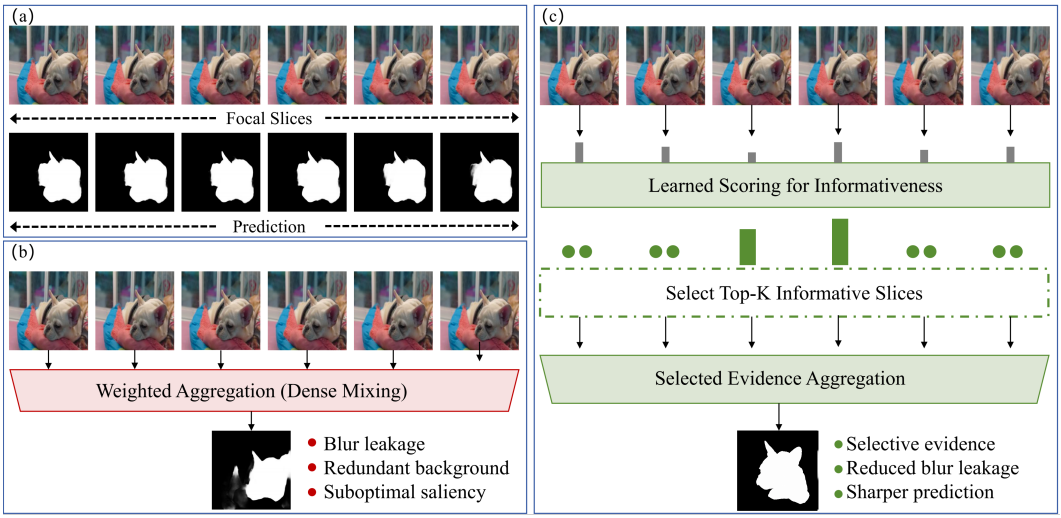


Figure 1: **Motivation for selective focal evidence aggregation.** (a) Focal slices and their corresponding saliency responses. (b) Dense weighted aggregation incorporates all slices, including potentially redundant or ambiguous observations. (c) Learned slice-wise scoring selects a top-K subset before aggregation. This schematic comparison motivates moving beyond individual informativeness to assess the conditional and joint predictive utility of focal observations. The displayed saliency maps are illustrative.

idence has been acquired. Complementarity requires looking beyond immediate individual gains. Even an adaptive selector can miss two slices whose joint benefit justifies their encoding cost when neither is worthwhile on its own. Stopping solely on the basis of single-slice gains can therefore leave useful combinations unexplored. The challenge is to evaluate both individual observations and their combinations in the context of the evidence already selected.

We introduce Complementarity-Aware Focal Selection and Aggregation (CFSA) to address these challenges. Building on loss-improvement supervision for dynamic feature selection Gadgil et al. (2024), CFSA learns the conditional prediction gains of candidate slices and slice pairs, with gains defined by reductions in the downstream saliency loss. Its pairwise look-ahead identifies combinations whose joint value may be underestimated by individual scoring. These estimates guide progressive evidence selection and aggregation, with adaptive stopping under an encoding budget. Both which slices are selected and when selection stops thus depend on their estimated contribution to the final prediction.

The value of selected evidence also depends on how it is combined with the all-focus representation. Both streams describe the same scene, sharing some structures while emphasizing different aspects of foreground and background. Their contributions to a common representation can therefore overlap. Accounting for this overlap requires considering how the two observations jointly constrain the representation. We connect focal selection with a shared-state formulation of fusion, in which cross-modal relations determine how the selected focal evidence and all-focus features jointly correct a spatial state.

Our Relation-Conditioned State-Space Fusion (RCSF) draws on the online-regression perspective of state-space models Liu et al. (2025). It treats all-focus and focal features as simultaneous observations, with their relationship determining the observation mappings and weights in a regularized joint correction. The resulting closed-form update couples the observation residuals through a common system that accounts for overlap between observation directions. Each observation's contribution consequently depends on the other, and this coupling is retained during subsequent spatial propagation. Together, CFSA and RCSF make complementarity a criterion for both choosing focal evidence and incorporating it into the saliency representation.

Our contributions are threefold:

- We formulate focal selection through conditional predictive utility and introduce CFSA, which combines single-slice and pairwise gain estimation with progressive aggregation and adaptive stopping to identify complementary focal evidence under an encoding budget.
- We develop RCSF, which uses cross-modal relations to parameterize a joint state correction that accounts for overlapping observation directions and preserves the resulting coupling during spatial propagation.
- Extensive experiments on four public LFSOD benchmarks demonstrate that the proposed network outperforms state-of-the-art approaches.

2 Related Work

2D and 3D Salient Object Detection Methods. Recent RGB-based SOD methods emphasize object integrity, fine structures, and semantic context. ICON Zhuge et al. (2023) and SelfReformer Yun & Lin (2024) address object completeness and structural preservation, while MENet Wang et al. (2023b) enhances representations at the pixel, region, and object levels. VSCode Luo et al. (2024) explores visual prompting for generalized salient and camouflaged object detection, and Samba He et al. (2025) introduces state-space modeling for long-range context. RGB-D methods complement appearance with geometric information, making cross-modal interaction central to their design. CIR-Net Cong et al. (2022) strengthens RGB–depth interaction, and HiDAnet Wu et al. (2023) develops hierarchical depth-aware representations. DCTNet Chen et al. (2024a) and MambaSOD Zhan et al. (2025) further investigate cross-modal fusion through disentangled Transformers and dual Mamba architectures, respectively. Despite these advances, foreground–background ambiguities can persist when appearance and depth provide insufficient discrimination. The focus variations and angular observations available in 4D light fields offer additional evidence for resolving these ambiguities, motivating saliency detection beyond RGB and RGB-D inputs.

Light Field Salient Object Detection Methods. Light-field SOD exploits appearance, focus, and parallax cues available from all-focus images, focal stacks, and sub-aperture views. Early work established the value of light-field observations for salient object localization Li et al. (2014), followed by a systematic review and benchmark Fu et al. (2022). Deep approaches introduced memory-oriented decoding in MoLF Zhang et al. (2019), learned light-field representations in Wang et al. Wang et al. (2019), and asymmetric two-stream processing in ERNet Piao et al. (2020b). Subsequent studies strengthened evidence integration through synergistic attention in SANet Zhang et al. (2021), edge-aware multimodal modeling in MEANet Jiang et al. (2022), and guided focal-stack refinement in GFRNet Yuan et al. (2023). More recent approaches explore adaptive fusion in LF-TransNet Liu et al. (2023), parallax-aware modeling by Yuan et al. Yuan et al. (2024), Transformer-based interaction in TLFNet Jiang et al. (2024), and SAM–Mamba integration in LFSamba Liu et al. (2024). These developments improve the extraction and interaction of light-field features. However, stronger feature interaction does not by itself determine which additional observations merit encoding given the evidence already acquired. Individually informative slices may repeat existing evidence, whereas individually weak slices may become useful together. CFSA addresses this selection problem by estimating conditional single-slice and pairwise prediction gains, guiding progressive acquisition and aggregation under an encoding budget. RCSF complements this design by treating all-focus and selected focal features as joint observations of a shared spatial state. Cross-modal relations govern their joint correction, accounting for overlap between observation directions as the fused state propagates spatially.

3 Method

Our network connects focal evidence selection with cross-modal fusion, as illustrated in Figure 2. CFSA progressively selects and aggregates focal observations according to their conditional predictive gains under an encoding budget, while RCSF integrates the selected evidence with all-focus features through relation-conditioned joint state correction.

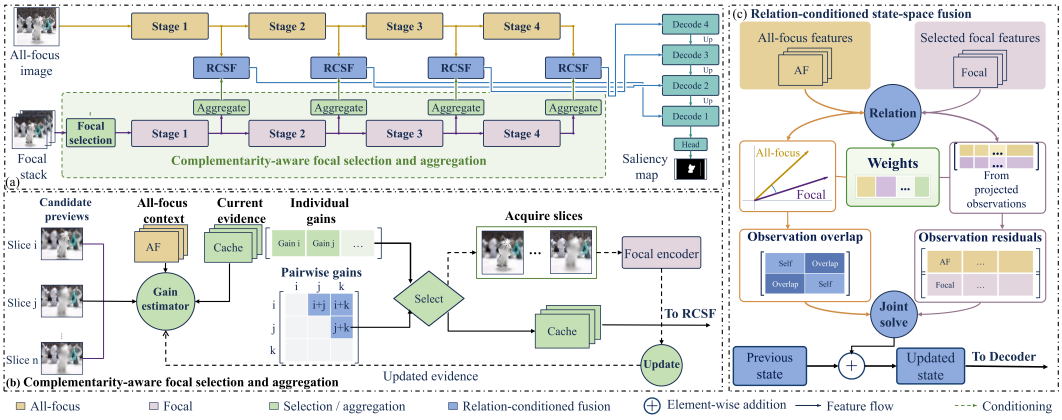


Figure 2: Overview of the proposed framework. **(a)** Overall architecture: all-focus and selected focal features are fused at multiple scales for saliency prediction. **(b)** CFSA estimates conditional gains of individual slices and slice pairs using candidate previews, all-focus context, and accumulated evidence, guiding progressive focal selection and aggregation. **(c)** RCSF uses cross-modal relations to determine observation mappings and weights, jointly correcting a shared spatial state while accounting for overlap between observation directions.

3.1 Complementarity-Aware Focal Selection and Aggregation

CFSA maintains an evolving focal representation and uses it to decide which observations to encode next. Let S denote the selected slice indices, $\mathbf{a}$ the cached all-focus features, and $\mathbf{z}_S$ the aggregated features of the selected slices. The current saliency prediction is $\widehat{Y}_S = P_\psi(\mathbf{a}, \mathbf{z}_S)$, where P_ψ comprises RCSF and the decoder. For an additional slice or slice pair A, we define its conditional gain using the prediction loss before and after acquisition:

$$\Delta(A \mid S) = \mathcal{L}_{\text{sal}}(\widehat{Y}_S, Y) - \mathcal{L}_{\text{sal}}(\widehat{Y}_{S\cup A}, Y), \tag{1}$$

where Y is the training mask. This definition ties selection to unresolved prediction errors. A slice that repeats already sufficient foreground evidence may have little gain, whereas one that helps resolve an ambiguous boundary can remain useful. Its value is therefore reassessed after each change to S.

The gain estimator receives lightweight previews of the focal stack, $\mathbf{a}$, and $\mathbf{z}_S$. The all-focus features provide scene context, while the accumulated focal representation supplies the evidence against which a candidate's additional value is assessed. For a pair, the estimator considers both candidate previews together and learns their joint gain from the prediction obtained after adding both slices. This exposes combinations that singleton decisions can miss: neither slice may resolve an ambiguity alone, yet their combined focus cues may support a better foreground–background separation.

At each step, CFSA compares single-slice and pairwise actions under the same acquisition rule:

$$A^* = \arg\max_{A\in\mathcal{C}_B(S)} \left[\widehat{\Delta}_\theta(A \mid S) - \lambda|A|\right], \tag{2}$$

where $\mathcal{C}_B(S)$ contains unselected singletons and unordered pairs satisfying $|S| + |A| \le B$, and $\lambda \ge 0$ is the per-slice encoding penalty. Selection stops if the candidate set is empty or the best score is nonpositive. Consequently, a pair can justify further acquisition even when both estimated singleton gains fall below the stopping threshold. The budget B bounds the number of slices passed through the focal encoder; candidate gains are estimated from previews.

After accepting A^*, CFSA updates $S \leftarrow S \cup A^*$, and only the newly acquired slices enter the focal encoder. Their features update $\mathbf{z}_S$ at each scale, and the resulting evidence conditions the next round of gain estimation. Cached features and incremental aggregation avoid re-encoding earlier selections. Aggregation is invariant to acquisition order, so selecting the same subset through different singleton or pairwise actions yields the same focal representation.

Following loss-improvement supervision Gadgil et al. (2024), we first train the saliency model with varying focal subsets, including an empty subset represented by $\mathbf{z}_\varnothing$. We then freeze this model and

train the gain estimator to regress Equation equation 1 over sampled subsets and actions. Holding the predictor fixed keeps the supervision tied to a consistent measure of utility. At inference, selection uses estimated gains and cached representations; RCSF and the decoder run once on the final subset.

3.2 Relation-Conditioned State-Space Fusion

RCSF determines how the selected focal evidence and all-focus features modify a common spatial representation. Their residuals indicate what the current state fails to explain, but the effect of each residual also depends on its observation direction. When two directions overlap, the corresponding corrections act on shared state components. RCSF resolves their contributions jointly, allowing each modality's correction to depend on evidence from the other.

At one feature scale, let t index positions along a spatial scan and $H_{t-1} \in \mathbb{R}^{d_s \times d_v}$ summarize preceding positions. The current all-focus and focal tokens are projected to values $\mathbf{v}_{a,t}, \mathbf{v}_{f,t} \in \mathbb{R}^{d_v}$. A relation encoder processes both tokens to generate observation directions $\mathbf{u}_{a,t}, \mathbf{u}_{f,t} \in \mathbb{R}^{d_s}$ and nonnegative weights $w_{a,t}, w_{f,t}$. Each direction specifies which linear combination of the state should explain its modality's value; its weight controls the strength of this constraint. Collecting the directions as columns of U_t, the values as rows of V_t, and the weights in diagonal matrix W_t, we obtain the observation model $V_t \approx U_t^\top H_t$.

Building on the online-regression perspective of Longhorn Liu et al. (2025), we determine the new state by jointly fitting both observations while retaining accumulated spatial information:

$$H_t = \arg\min_H \left\{ \frac{\gamma}{2}\|H - H_{t-1}\|_F^2 + \frac{1}{2}\left\|W_t^{1/2}(U_t^\top H - V_t)\right\|_F^2 \right\}, \tag{3}$$

where $\gamma > 0$ regulates changes to the preceding state. Defining $O_t = U_t W_t^{1/2}$ and $\widetilde{V}_t = W_t^{1/2} V_t$, the solution is

$$\begin{aligned} H_t &= H_{t-1} + O_t G_t^{-1}\left(\widetilde{V}_t - O_t^\top H_{t-1}\right), \\ G_t &= \gamma I_2 + O_t^\top O_t. \end{aligned} \tag{4}$$

The bracketed term contains the two observation residuals. Their correction coefficients are determined together by the 2×2 system G_t, whose off-diagonal entry is $\sqrt{w_{a,t} w_{f,t}}\, \mathbf{u}_{a,t}^\top \mathbf{u}_{f,t}$. For orthogonal observation directions, the residual corrections decouple. With overlapping directions, the all-focus correction coefficient depends on the focal residual and vice versa. Thus, the relation determines both where each observation modifies the state and how that modification interacts with the other observation. Positive regularization makes this joint correction well-defined even when the directions coincide.

The corrected state is carried to the next spatial position, so this interaction also influences subsequent evidence integration. Because the observation parameters depend only on encoded tokens, Equation equation 4 is affine in H_{t-1} and admits recurrent or parallel prefix evaluation. Compact two-row heads ($d_s = 2$) preserve the full transition, including its off-diagonal interactions, starting from $H_0 = 0$. Each head is read out through an input-conditioned query, $\mathbf{o}_t = H_t^\top \mathbf{q}_t$, and the head outputs form the fused features supplied to the saliency decoder.

4 Experiments

4.1 Experimental Setup

Datasets and evaluation metrics. We evaluate on four light-field salient object detection benchmarks: DUTLF-FS (Wang et al., 2019), HFUT (Zhang et al., 2017), LFSD (Li et al., 2014), and DUTLF-V2 (Piao et al., 2020a). The DUTLF-FS protocol uses 1,000 training samples and a held-out test partition. HFUT and LFSD contain 255 and 100 samples, respectively, while DUTLF-V2 contains 4,204 samples in total. We report four evaluation metrics: mean absolute error (MAE), structure-measure (S_α) (Fan et al., 2017), enhanced-alignment measure (E_S) (Fan et al., 2018), and weighted F-measure (F_β^w) (Margolin et al., 2014). Lower MAE and higher values of the remaining metrics indicate better performance. We also report precision–recall curves to assess predictions across decision thresholds.

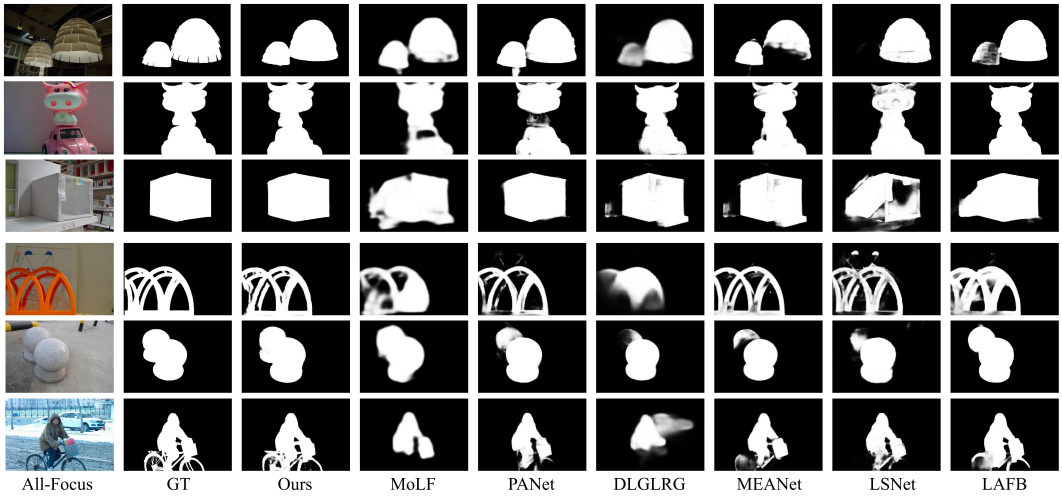


Figure 3: Qualitative comparison on six scenes. From left to right: all-focus image, ground truth, ours, MoLF, PANet, DLGLRG, MEANet, LSNet, and LAFB. The examples compare foreground completeness, preservation of fine structures, and separation from the background.

Implementation details. Our model is implemented in PyTorch and trained on four NVIDIA GeForce RTX 3080 GPUs. We use Adam with an initial learning rate of 5×10^{-4}, weight decay of 10^{-4}, batch size 4, and a 200-epoch training schedule. Model checkpoints are selected by validation performance. We first train the saliency predictor and then optimize CFSA's conditional-gain estimator, keeping the encoders, focal aggregation component, RCSF, and decoder fixed during the second stage. At inference, CFSA reuses lightweight previews and cached features to select and aggregate focal evidence. RCSF and the decoder run once after selection is complete.

4.2 Comparison with SOTA Methods

We compare our method with 37 SOD baselines, comprising 22 light-field methods, seven RGB-only methods, and eight two-modality methods. The light-field baselines include two traditional approaches, WSC (Li et al., 2015) and LFS (Li et al., 2017), and twenty deep-learning-based methods: MoLF (Zhang et al., 2019), LFNet (Zhang et al., 2020), ERNet (Piao et al., 2020b), PANet (Piao et al., 2023), DLGLRG (Liu et al., 2021), OBGNet (Jing et al., 2021), SANet (Zhang et al., 2021), MEANet (Jiang et al., 2022), CDINet (Chen et al., 2024b), GFRNet (Yuan et al., 2023), SRINet (Lv et al., 2023), LFTransNet (Liu et al., 2023), IRNet (Liao & Gao, 2024), TLFNet (Jiang et al., 2024), LFSODNet (Zheng et al., 2024), FEANet (Huang et al., 2024), VMKNet (Han et al., 2025), LF-Mamba (Geng et al., 2025), MMCFNet (Hu et al., 2025), and CCNet (Zhang et al., 2026).

The RGB-only baselines are EDN (Wu et al., 2022), BBRF (Ma et al., 2023), LeNo (Wang et al., 2023a), CTSOD (Sun et al., 2025), MDSAM (Gao et al., 2024), STDNet (Su et al., 2025), and UGRAN (Yuan et al., 2025). The two-modality baselines are SwinNet (Liu et al., 2022), PICR-Net (Cong et al., 2023), LSNet (Zhou et al., 2023), HRTransNet (Tang et al., 2023), CPNet (Hu et al., 2024), LAFB (Wang et al., 2024), DiffSOD (Song et al., 2025), and ADENet (Yi et al., 2025).

For the compared methods, we run the original authors' released code or use their publicly released predictions and evaluation results. All results for our method and its ablation variants are obtained from our experiments. RGB-only methods take all-focus images as input, while two-modality methods take all-focus images paired with depth. Architectures originally developed for RGB-T detection therefore use depth as the auxiliary input in this comparison. Light-field methods use their required focal-stack or multi-view inputs. Tables 1 and 2 report the light-field and RGB/two-modality comparisons, respectively.

Quantitative evaluation. Table 1 shows that our method achieves the lowest MAE and the highest S_α, E_S, and F_β^w on all four benchmarks. The agreement across these criteria indicates that reduced pixelwise error is accompanied by better structural and foreground prediction quality. The strongest

Table 1: Comparison with light-field SOD methods on four benchmarks. T and D-L denote traditional and deep-learning-based methods. The best and second-best results in each column are shown in **bold** and with underlining, respectively. Ties are determined at the displayed precision; "–" denotes an unavailable result.

Type	Method	Venue/Year	DUTLF-FS				HFUT				LFSD				DUTLF-V2			
			MAE↓	$S_\alpha\uparrow$	$E_S\uparrow$	$F_\beta^w\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_S\uparrow$	$F_\beta^w\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_S\uparrow$	$F_\beta^w\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_S\uparrow$	$F_\beta^w\uparrow$
T	WSC	CVPR15	.157	.649	.780	.593	.154	.620	.620	.495	.152	.770	.697	.717	.167	.726	.614	.486
	LFS	TPAMI17	.238	.579	.694	.365	.212	.581	.693	.345	.212	.674	.580	.510	.217	.578	.527	.327
D-L	MoLF	NeurIPS19	.051	.866	.923	.843	.076	.786	.826	.671	.087	.833	.888	.815	.064	.825	.866	.724
	LFNet	TIP20	.055	.878	.913	.833	.096	.736	.770	.620	.092	.820	.882	.805	.073	.817	.852	.716
	ERNet	AAAI20	.039	.899	.947	.908	.084	.778	.843	.722	.081	.834	.887	.849	.074	.839	.891	.851
	PANet	TCYB23	.039	.907	.944	.897	.074	.794	.853	.724	.076	.848	.892	.848	.073	.841	.887	.849
	DLGLRG	ICCV21	.030	.928	.958	.924	.071	.763	.841	.708	.069	.866	.903	.861	.046	.860	.908	.815
	OBGNet	ACM MM21	–	–	–	–	.092	.747	.799	.664	.078	.852	.897	.858	.037	.896	.937	.866
	SANet	BMVC21	.032	.918	.954	.919	.078	.784	.855	.734	.073	.841	.889	.844	.065	.866	.903	.834
	MEANet	IJON22	.066	.853	.919	.862	.075	.851	.891	.849	.084	.758	.811	.674	.056	.843	.895	.784
	CDINet	TCSVT24	–	–	–	–	–	–	–	–	.047	.880	.908	.826	.032	.905	.940	.879
	GFRNet	ICME23	.027	.931	.962	.932	.068	.868	.892	.866	.067	.861	.907	.845	.034	.891	.922	.861
	SRINet	JVCI23	.030	.927	.958	.927	.066	.780	.844	.780	.067	.862	.900	.861	.052	.887	.936	.875
	LFTransNet	TCSVT23	.023	.938	.971	.950	.067	.835	.886	.804	.048	.900	.933	.902	.028	.899	.942	.894
	IRNet	TOMM24	.026	.933	.965	.932	.060	.877	.916	.878	.061	.838	.876	.776	.047	.894	.945	.897
	TLFNet	TIP24	.026	.933	.964	.933	.062	.876	.915	.881	.058	.842	.886	.793	.038	.890	.933	.859
	LFSODNet	TCSVT24	.022	.943	.963	.943	.066	.835	.871	.784	.053	.889	.911	.878	.029	.923	.908	.890
	FEANet	ICME24	.030	.927	.958	.936	.074	.798	.852	.750	.071	.855	.897	.853	.068	.861	.929	.825
	VMKNet	TMM25	.018	.944	.975	.955	.060	.838	.888	.811	.047	.892	.932	.900	.029	.909	.948	.892
	LFMamba	ACM MM25	.018	.951	.978	.958	.048	.903	.921	.891	.051	.896	.929	.895	.026	.915	.951	.899
	MMCFNet	IVC25	.023	.942	.964	.942	.064	.838	.875	.787	.045	.905	.922	.904	.037	.886	.927	.862
	CCNet	JEI26	.026	.935	.959	.926	.069	.819	.874	.764	.055	.892	.918	.887	.037	.885	.932	.889
D-L	Ours	2026	**.014**	**.962**	**.982**	**.963**	**.044**	**.906**	**.922**	**.904**	**.042**	**.941**	**.936**	**.909**	**.022**	**.927**	**.966**	**.921**

Table 2: Comparison with RGB-only (2D) and two-modality (3D) SOD methods. Our light-field method is listed as the 4D model. The best and second-best results in each column are shown in **bold** and with underlining, respectively. Ties are determined at the displayed precision.

Type	Method	Venue/Year	DUTLF-FS				HFUT				LFSD				DUTLF-V2			
			MAE↓	$S_\alpha\uparrow$	$E_S\uparrow$	$F_\beta^w\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_S\uparrow$	$F_\beta^w\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_S\uparrow$	$F_\beta^w\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_S\uparrow$	$F_\beta^w\uparrow$
2D	EDN	TIP22	.043	.904	.939	.897	.081	.806	.835	.726	.076	.853	.897	.859	.034	.903	.936	.861
	BBRF	TIP23	.030	.927	.959	.929	.076	.804	.853	.762	.088	.811	.847	.768	.051	.869	.907	.843
	LeNo	AAAI23	.054	.896	.904	.826	.082	.709	.801	.698	.068	.869	.902	.873	.047	.869	.922	.868
	CTSOD	TPAMI25	.047	.889	.935	.889	.087	.746	.825	.698	.111	.779	.835	.778	.057	.847	.903	.806
	MDSAM	ACM MM24	.031	.930	.954	.932	.072	.833	.874	.789	.066	.879	.918	.889	.044	.905	.940	.898
	STDNet	TPAMI25	.029	.932	.957	.933	.069	.838	.881	.795	.063	.883	.921	.891	.041	.908	.943	.901
	UGRAN	TIP25	.050	.906	.942	.911	.089	.792	.835	.732	.090	.850	.885	.841	.062	.795	.841	.869
3D	SwinNet	TCSVT22	.035	.916	.962	.925	.063	.842	.876	.805	.068	.867	.914	.871	.047	.872	.883	.867
	PICRNet	ACM MM23	.021	.943	.971	.946	.078	.798	.851	.745	.052	.889	.923	.869	.046	.872	.921	.889
	LSNet	TIP23	.062	.868	.909	.859	.093	.743	.810	.682	.085	.834	.878	.852	.086	.764	.857	.707
	HRTransNet	TCSVT23	.024	.935	.963	.934	.058	.883	.921	.887	.064	.830	.865	.769	.048	.872	.916	.828
	CPNet	IJCV24	.022	.940	.971	.937	.068	.818	.863	.776	.074	.850	.899	.854	.029	.911	.945	.891
	LAFB	TCSVT24	.027	.930	.961	.927	.074	.799	.845	.736	.073	.849	.903	.848	.038	.888	.931	.862
	DiffSOD	TIP25	.020	.939	.959	.941	.060	.872	.914	.875	.057	.875	.914	.865	.028	.913	.947	.894
	ADENet	SPL25	.027	.933	.963	.948	.075	.811	.864	.798	.069	.865	.906	.891	.037	.883	.931	.879
4D	Ours	2026	**.014**	**.962**	**.982**	**.963**	**.044**	**.906**	**.922**	**.904**	**.042**	**.941**	**.936**	**.909**	**.022**	**.927**	**.966**	**.921**

competing methods vary across datasets and metrics, whereas our model maintains the leading results throughout. This consistency supports the effectiveness of the complete framework across the evaluated light-field benchmarks.

The improvements are not uniform across criteria. On HFUT, alignment scores are close to those of the strongest baseline, while the weighted foreground score shows a clearer separation; on LFSD, the structural improvement is more pronounced. These differences show why evaluating several aspects of prediction quality is useful. Comparisons with VMKNet and LFMamba are also relevant: these methods already exploit focal selection or state-space modeling. The gains over these baselines support the proposed combination of conditional focal selection and relation-conditioned fusion at the framework level.

Table 2 extends the comparison to RGB-only and two-modality architectures, including models using depth as auxiliary input. Our method leads on all reported criteria, placing it favorably against both appearance-based and depth-assisted alternatives. Together, the two tables establish the com-

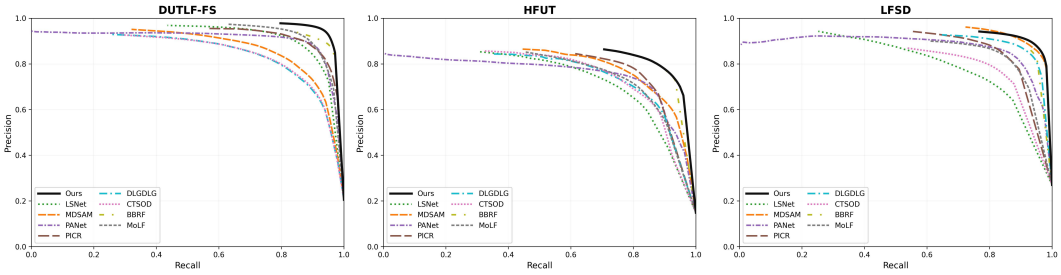


Figure 4: Precision–recall curves of our method and representative competing methods on DUTLF-FS, HFUT, and LFSD.

Table 3: Ablation study of CFSA and RCSF. The reference variant (w/o CFSA & RCSF) densely encodes the focal stack and fuses all-focus and focal features by concatenation. The w/o CFSA variant removes CFSA while retaining RCSF. The best result in each column is marked in **bold**.

Variant	DUTLF-FS				HFUT				LFSD				DUTLF-V2			
	MAE↓	$S_\alpha\uparrow$	$E_S\uparrow$	$F_\beta^w\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_S\uparrow$	$F_\beta^w\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_S\uparrow$	$F_\beta^w\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_S\uparrow$	$F_\beta^w\uparrow$
w/o CFSA & RCSF	.029	.928	.958	.928	.067	.871	.894	.866	.062	.871	.910	.870	.038	.899	.938	.881
w/o CFSA	.022	.941	.968	.929	.054	.872	.910	.891	.067	.872	.919	.881	.029	.914	.942	.886
Ours	**.014**	**.962**	**.982**	**.963**	**.044**	**.906**	**.922**	**.904**	**.042**	**.941**	**.936**	**.909**	**.022**	**.927**	**.966**	**.921**

petitiveness of the complete light-field model against general saliency architectures and methods designed specifically for focal observations.

Figure 4 complements these scalar comparisons by examining predictions across decision thresholds. Our curves remain near the upper envelope over much of the plotted range, with a clear advantage on HFUT. This behavior indicates that favorable foreground recovery extends across a range of precision–recall trade-offs, rather than being confined to a single threshold.

Qualitative evaluation. Figure 3 illustrates how these improvements appear in the predicted saliency maps. In scenes with multiple foreground components, several competing methods omit target regions or merge adjacent components. Our predictions recover more complete objects while retaining the separation between distinct regions. The bicycle example further shows better preservation of narrow, connected structures, and our predictions also maintain coherent foreground interiors with fewer missing regions.

Recovering an object's interior while preserving thin structures and intervening background regions requires both foreground coverage and spatial discrimination. The examples show that our model can maintain these properties together, consistent with the improvements in structural and weighted foreground scores. Differences are smaller for compact objects with simple boundaries, where several methods already approximate the annotations closely. The more informative comparisons therefore concern the preservation of object structure, beyond identifying the approximate location of a salient region.

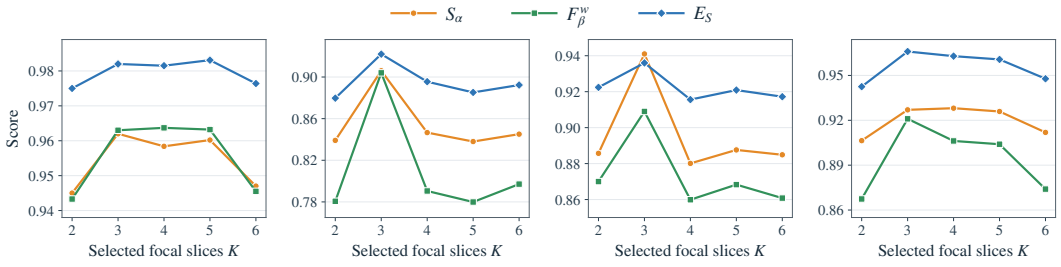
4.3 Ablation Study

The ablations examine how the treatment of focal evidence affects prediction quality. We first compare the complete selection–fusion design with dense focal encoding and concatenation-based fusion, then vary the number of selected slices to assess whether additional observations consistently improve the prediction.

Joint contribution of CFSA and RCSF. Table 3 compares the full model with a reference variant that densely encodes the focal stack and concatenates all-focus and focal features. Both configurations receive the same input modalities, and the reference already has access to the complete stack. Nevertheless, the proposed configuration improves every reported metric across all four benchmarks. Access to more focal observations is therefore insufficient to match the full model under this reference fusion scheme.

Table 4: Effect of the fixed number of selected focal slices K. The best result in each column is marked in **bold**. Values are rounded to three decimal places; rankings use the available unrounded values.

K	DUTLF-FS				HFUT				LFSD				DUTLF-V2			
	MAE↓	$S_\alpha\uparrow$	$E_S\uparrow$	$F_\beta^w\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_S\uparrow$	$F_\beta^w\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_S\uparrow$	$F_\beta^w\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_S\uparrow$	$F_\beta^w\uparrow$
2	.019	.945	.975	.943	.065	.839	.880	.781	.054	.886	.922	.870	.032	.906	.943	.867
3	**.014**	**.962**	.982	.963	**.044**	**.906**	**.922**	**.904**	**.042**	**.941**	**.936**	**.909**	**.022**	.927	**.966**	**.921**
4	.015	.958	.982	**.964**	.053	.847	.896	.791	.056	.880	.916	.860	.023	**.928**	.963	.906
5	.015	.960	**.983**	.963	.056	.838	.885	.780	.052	.888	.921	.868	.024	.926	.961	.904
6	.019	.947	.976	.946	.057	.845	.892	.797	.053	.885	.917	.861	.029	.912	.948	.874



(a) DUTLF-FS (b) HFUT (c) LFSD (d) DUTLF-V2

Figure 5: Effect of the fixed number of selected focal slices on four benchmarks. Curves show S_α, F_β^w, and $E_S\uparrow$ from Table 4. Higher values indicate better performance. Each panel uses its own vertical scale; MAE is reported in the table.

The gains span pixelwise error, structural agreement, and foreground quality, supporting the combined design for selecting and integrating focal evidence. CFSA assesses candidate observations in the context of accumulated evidence, while RCSF uses cross-modal relations to determine their joint contribution to a spatial state. Since both modules change together, this comparison measures their combined contribution rather than isolating the effect of either module.

Number of selected focal slices. Table 4 and Figure 5 compare fixed subset sizes $K \in \{2, 3, 4, 5, 6\}$. Increasing K from two to three improves every reported metric on all four benchmarks, indicating that an overly restricted subset can leave useful focal evidence unexploited. Further increases do not yield consistent gains. Within the tested range, $K = 3$ provides the strongest overall results, leading on all criteria on HFUT and LFSD and on most criteria on DUTLF-V2.

The exceptions clarify the role of subset size. On DUTLF-FS, E-measure peaks at $K = 5$ and weighted F-measure at $K = 4$, while structural agreement on DUTLF-V2 is slightly higher at $K = 4$. Larger subsets can therefore benefit individual aspects of prediction without improving the result across metrics. Figure 5 also shows distinct responses across benchmarks: DUTLF-FS changes relatively little between three and five slices, whereas HFUT and LFSD favor the more compact configuration. These observations support evaluating additional focal evidence by its predictive contribution. Selecting too few slices can limit the available information, but allocating more encoding capacity does not guarantee a better representation.

5 Conclusion

We presented a complementarity-aware network for light-field salient object detection. CFSA guides budgeted focal selection through conditional gains of individual slices and slice pairs, while RCSF integrates all-focus and focal evidence through relation-conditioned joint state correction. Experiments on four benchmarks demonstrate the effectiveness of the network, and ablations show that additional focal observations do not consistently improve prediction. These results motivate selecting focal evidence according to its predictive contribution.

AI USE STATEMENT

We used generative AI tools to assist with writing and polishing the manuscript, including refining wording, correcting grammar, and improving clarity and readability. We also used these tools to support literature retrieval and discovery, including identifying potentially relevant prior work and locating bibliographic information. The authors take full responsibility for the final content of this paper, including its scientific claims, experimental results, and the accuracy of its references.

REFERENCES

Hao Chen, Feihong Shen, Ding Ding, Yongjian Deng, and Chao Li. Disentangled cross-modal transformer for RGB-D salient object detection and beyond. *IEEE Transactions on Image Processing*, 33:1699–1709, 2024a. doi: 10.1109/TIP.2024.3364022.

Yilei Chen, Gongyang Li, Ping An, Zhi Liu, X. Huang, and Qiang Wu. Light field salient object detection with sparse views via complementary and discriminative interaction network. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(2):1070–1085, 2024b. doi: 10.1109/TCSVT.2023.3290600.

Runmin Cong, Qinwei Lin, Chen Zhang, Chongyi Li, Xiaochun Cao, Qingming Huang, and Yao Zhao. CIR-Net: Cross-modality interaction and refinement for RGB-D salient object detection. *IEEE Transactions on Image Processing*, 31:6800–6815, 2022. doi: 10.1109/TIP.2022.3216198.

Runmin Cong, Hongyu Liu, Chen Zhang, Wei Zhang, Feng Zheng, Ran Song, and Sam Kwong. Point-aware interaction and CNN-induced refinement network for RGB-D salient object detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 406–416, 2023.

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*, pp. 4548–4557, 2017. URL https://arxiv.org/abs/1708.00786.

Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 698–704, 2018. doi: 10.24963/ijcai.2018/97.

Keren Fu, Yao Jiang, Ge-Peng Ji, Tao Zhou, Qijun Zhao, and Deng-Ping Fan. Light field salient object detection: A review and benchmark. *Computational Visual Media*, 8(4):509–534, 2022. doi: 10.1007/s41095-021-0256-2.

Soham Gadgil, Ian Covert, and Su-In Lee. Estimating conditional mutual information for dynamic feature selection. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Oju2Qu9jvn.

Shixuan Gao, Pingping Zhang, Tianyu Yan, and Huchuan Lu. Multi-scale and detail-enhanced segment anything model for salient object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. URL https://arxiv.org/abs/2408.04326.

Xinbo Geng, Fan Shi, Xu Cheng, Chen Jia, Meng Zhao, and Shengyong Chen. LFMamba: Focal stack-aware state space modeling for light field salient object detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2074–2083, 2025. doi: 10.1145/3746027.3755654.

Jiaxin Han, Feng Li, Anqi Li, Mengmeng Zhang, Huihui Bai, Jimin Xiao, and Yao Zhao. Enhancing light field salient object detection with variance-maximized key focal slice selection. *IEEE Transactions on Multimedia*, 27:6555–6567, 2025. doi: 10.1109/TMM.2025.3586131.

Jiahao He, Keren Fu, Xiaohong Liu, and Qijun Zhao. Samba: A unified Mamba-based framework for general salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25314–25324, 2025. doi: 10.1109/CVPR52734.2025.02357.

Xihang Hu, Fuming Sun, Jing Sun, Fasheng Wang, and Haojie Li. Cross-modal fusion and progressive decoding network for RGB-D salient object detection. *International Journal of Computer Vision*, 132(8):3067–3085, 2024. doi: 10.1007/s11263-024-02020-y.

Xin Hu, Fen Chen, Zongju Peng, Lian Huang, and Jiawei Xu. MMCFNet: Multi-scale and multimodal complementary fusion network for light field salient object detection. *Image and Vision Computing*, 162:105680, 2025. URL https://www.sciencedirect.com/science/article/abs/pii/S0262885625002689.

Ziyi Huang, Binbin Yan, Shuo Chen, Dongliang Wang, and Lu Yang. Focal stack alignment enhancement network for light field salient object detection. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1–6. IEEE, 2024. doi: 10.1109/ICME57554.2024.10687985.

Yao Jiang, Wenbo Zhang, Keren Fu, and Qijun Zhao. MEANet: Multi-modal edge-aware network for light field salient object detection. *Neurocomputing*, 491:78–90, 2022. doi: 10.1016/j.neucom.2022.03.056.

Yao Jiang, Xin Li, Keren Fu, and Qijun Zhao. Transformer-based light field salient object detection and its application to autofocus. *IEEE Transactions on Image Processing*, 33:6647–6659, 2024. doi: 10.1109/TIP.2024.3498331.

Dong Jing, S. Zhang, Runmin Cong, and Y. Lin. Occlusion-aware bi-directional guided network for light field salient object detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. doi: 10.1145/3474085.3475312.

Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2806–2813, 2014.

Nianyi Li, Bilin Sun, and Jingyi Yu. A weighted sparse coding framework for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5216–5223, 2015. doi: 10.1109/CVPR.2015.7299158.

Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1605–1616, 2017.

Guibiao Liao and Wei Gao. Rethinking feature mining for light field salient object detection. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2024. doi: 10.1145/3676967.

Bo Liu, Rui Wang, Lemeng Wu, Yihao Feng, Peter Stone, and Qiang Liu. Longhorn: State space models are amortized online learners. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8jOqCcLzeO.

Nian Liu, Wangbo Zhao, Dingwen Zhang, Junwei Han, and Ling Shao. Light field saliency detection with dual local graph learning and reciprocative guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2110.00698.

Zhengyi Liu, Yacheng Tan, Qian He, and Y. Xiao. SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4486–4497, 2022.

Zhengyi Liu, Qian He, Linbo Wang, Xianyong Fang, and Bin Tang. LFTransNet: Light field salient object detection via a learnable weight descriptor. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7764–7773, 2023. doi: 10.1109/TCSVT.2023.3281465.

Zhengyi Liu, Longzhen Wang, Xianyong Fang, Zhengzheng Tu, and Linbo Wang. LFSamba: Marry SAM with Mamba for light field salient object detection. *IEEE Signal Processing Letters*, 31:3144–3148, 2024. doi: 10.1109/LSP.2024.3493799.

Ziyang Luo, Nian Liu, Wangbo Zhao, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Khan, and Junwei Han. VSCode: General visual salient and camouflaged object detection with 2D prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17169–17180, 2024. doi: 10.1109/CVPR52733.2024.01625.

Chengtao Lv, Xiaofei Zhou, Bin Zhu, Deyang Liu, Bolun Zheng, Jiyong Zhang, and Chenggang Yan. SRI-Net: Similarity retrieval-based inference network for light field salient object detection. *Journal of Visual Communication and Image Representation*, 90:103721, 2023. doi: 10.1016/j.jvcir.2022.103721.

Mingcan Ma, Changqun Xia, Chenxi Xie, Xiaowu Chen, and Jia Li. Boosting broader receptive fields for salient object detection. *IEEE Transactions on Image Processing*, 32:1026–1038, 2023. doi: 10.1109/TIP.2022.3232209.

Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2014. doi: 10.1109/CVPR.2014.39.

Yongri Piao, Zhengkun Rong, Shuang Xu, Miao Zhang, and Huchuan Lu. DUT-LFSaliency: Versatile dataset and light field-to-RGB saliency detection, 2020a. URL https://arxiv.org/abs/2012.15124.

Yongri Piao, Zhengkun Rong, Miao Zhang, and Huchuan Lu. Exploit and replace: An asymmetrical two-stream architecture for versatile light field saliency detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11865–11873, 2020b. doi: 10.1609/aaai.v34i07.6860.

Yongri Piao, Y. Jiang, Miao Zhang, J. Wang, and Huchuan Lu. PANet: Patch-aware network for light field salient object detection. *IEEE Transactions on Cybernetics*, 2023.

Mengke Song, Luming Li, Xu Yu, and Chenglizhao Chen. Pushing the boundaries of salient object detection: A denoising-driven approach. *IEEE Transactions on Image Processing*, 34:3903–3917, 2025. doi: 10.1109/TIP.2025.3576993.

Zhuo Su, Li Liu, Matthias Müller, Jiehua Zhang, Diana Wofk, Ming-Ming Cheng, and Matti Pietikäinen. Rapid salient object detection with difference convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. URL https://ieeexplore.ieee.org/document/11072100.

Ke Sun, Zhongxi Chen, Xianming Lin, Xiaoshuai Sun, Hong Liu, and Rongrong Ji. Conditional diffusion models for camouflaged and salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2833–2848, 2025. doi: 10.1109/TPAMI.2025.3527469.

Bin Tang, Zhengyi Liu, Yacheng Tan, and Qian He. HRTransNet: HRFormer-driven two-modality salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2):728–742, 2023. URL https://arxiv.org/abs/2301.03036.

He Wang, Lin Wan, and He Tang. LeNo: Adversarial robust salient object detection networks with learnable noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2023a. URL https://arxiv.org/abs/2210.15392.

K. Wang, Zhengzheng Tu, Chenglong Li, C. Zhang, and Bin Luo. Learning adaptive fusion bank for multi-modal salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):7344–7358, 2024.

Tiantian Wang, Yongri Piao, Xiao Li, Lihe Zhang, and Huchuan Lu. Deep learning for light field saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8838–8848, 2019.

Yi Wang, Ruili Wang, Xin Fan, Tianzhu Wang, and Xiangjian He. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10031–10040, 2023b. doi: 10.1109/CVPR52729.2023.00967.

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. doi: 10.1109/TIP.2022.3164550.

Zongwei Wu, Guillaume Allibert, Fabrice Meriaudeau, Chao Ma, and Cédric Demonceaux. HiDAnet: RGB-D salient object detection via hierarchical depth awareness. *IEEE Transactions on Image Processing*, 32:2160–2173, 2023. doi: 10.1109/TIP.2023.3263111.

Kang Yi, Yumeng Li, Haoran Tang, and Jing Xu. Adaptive depth enhancement network for RGB-D salient object detection. *IEEE Signal Processing Letters*, 32:176–180, 2025. doi: 10.1109/LSP.2024.3506863.

Bo Yuan, Yao Jiang, Keren Fu, and Qijun Zhao. Guided focal stack refinement network for light field salient object detection. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 2387–2392, 2023. doi: 10.1109/ICME55011.2023.00407.

Bo Yuan, Yao Jiang, Keren Fu, and Qijun Zhao. Parallax-aware network for light field salient object detection. *IEEE Signal Processing Letters*, 31:810–814, 2024. doi: 10.1109/LSP.2024.3374079.

Yao Yuan, Pan Gao, Qun Dai, Jie Qin, and Wei Xiang. Uncertainty-guided refinement for fine-grained salient object detection. *IEEE Transactions on Image Processing*, 34:2301–2314, 2025. doi: 10.1109/TIP.2025.3557562.

Yi Ke Yun and Weisi Lin. Towards a complete and detail-preserved salient object detection. *IEEE Transactions on Multimedia*, 26:4667–4680, 2024. doi: 10.1109/TMM.2023.3325731.

Yue Zhan, Zhihong Zeng, Haijun Liu, Xiaoheng Tan, and Yinli Tian. MambaSOD: Dual Mamba-driven cross-modal fusion network for RGB-D salient object detection. *Neurocomputing*, 631: 129718, 2025. doi: 10.1016/j.neucom.2025.129718.

Jun Zhang, Meng Wang, Liang Lin, Xun Yang, Jun Gao, and Yong Rui. Saliency detection on light field: A multi-cue approach. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 13(3):32:1–32:22, 2017. doi: 10.1145/3107956.

Miao Zhang, Jingjing Li, Wei Ji, Yongri Piao, and Huchuan Lu. Memory-oriented decoder for light field salient object detection. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper/8376-memory-oriented-decoder-for-light-field-salient-object-detection.

Miao Zhang, Wei Ji, Yongri Piao, J. Li, Y. Zhang, Shuang Xu, and Huchuan Lu. LFNet: Light field fusion network for salient object detection. *IEEE Transactions on Image Processing*, 29: 6276–6287, 2020. URL https://ieeexplore.ieee.org/document/9082882.

Yi Zhang, Geng Chen, Qian Chen, Yujia Sun, Yong Xia, Olivier Deforges, Wassim Hamidouche, and Lu Zhang. Learning synergistic attention for light field salient object detection. In *Proceedings of the 32nd British Machine Vision Conference*, 2021. URL https://arxiv.org/abs/2104.13916.

Yunzuo Zhang, Shuangshuang Wang, Zhiwei Tu, and Yuehui Yang. Cross-modal co-attention network for light field salient object detection. *Journal of Electronic Imaging*, 35(1):013020, 2026. doi: 10.1117/1.JEI.35.1.013020.

Xin Zheng, Zhengqu Li, Deyang Liu, Xiaofei Zhou, and Caifeng Shan. Spatial attention-guided light field salient object detection network with implicit neural representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12):12437–12449, 2024. doi: 10.1109/TCSVT.2024.3437685.

Wujie Zhou, Y. Zhu, Jingsheng Lei, R. Yang, and Lu Yu. LSNet: Lightweight spatial boosting network for detecting salient objects in RGB-thermal images. *IEEE Transactions on Image Processing*, 32:1329–1340, 2023.

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, 2023. doi: 10.1109/TPAMI.2022.3179526.