LEARNING WHEN TO TRUST DEPTH: SCALE-WISE LATENT RELIABILITY ROUTING FOR RGB-D SALIENT OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Depth benefits RGB-D salient object detection only when its geometric evidence is reliable and relevant to the target; noisy, incomplete, or misaligned depth can instead cause negative cross-modal transfer. We formulate RGB-D fusion as inference of a *scale-wise latent reliability variable* that controls how the modalities interact at each representation level. Our framework, MPLoRA-MoE, combines shared semantic encoding, modality-specific low-rank adaptation, and four heterogeneous fusion experts. At each scale, an independent router predicts a soft expert distribution from pre-fusion RGB and depth features, allowing depth evidence to be preserved, attenuated, or reconciled without quality or profile annotations. The routed hierarchy is decoded through a single edge-semantic pathway, so conditional computation changes the fused representation rather than selecting among multiple outputs. Across eight RGB-D SOD benchmarks, MPLoRA-MoE achieves state-of-the-art performance on multiple datasets and remains competitive on the others. Ablations further show that the gains depend on reliability-oriented expert specialization, while moderate routing sharpness and compact modality-specific adaptation are sufficient. These results support learning when and how strongly depth should contribute instead of treating it as uniformly trustworthy. Code will be released upon publication.

1 INTRODUCTION

RGB-D salient object detection (SOD) localizes visually dominant objects by combining RGB appearance with depth geometry. Depth is valuable when RGB contrast is weak, illumination is poor, or texture distractors obscure boundaries (Peng et al., 2014; Fan et al., 2021; Bao et al., 2026). Its contribution, however, is conditional: depth from sensors, stereo pipelines, or monocular estimation may contain missing regions, quantization artifacts, shifted contours, or structures unrelated to the salient target. Injecting such cues indiscriminately can introduce false responses or distort boundaries, making multimodal prediction less reliable than RGB alone.

The key question is not whether a depth map is globally "good," but whether its geometry is useful *relative to RGB for the current target*. Figure 1(a) illustrates four reliability relations. When both modalities are informative, their cues should be fused complementarily; misleading depth should be suppressed when RGB is reliable, while useful geometry should be preserved when depth resolves RGB ambiguity. If both modalities are uncertain, fusion should remain flexible rather than commit prematurely to either source. These cases require qualitatively different fusion behaviors.

Existing RGB-D SOD methods have advanced cross-modal representation learning through improved encoding, cooperative fusion, depth calibration, modality-specific modeling, and Transformer-based interaction (Fan et al., 2020; Fu et al., 2020; Ji et al., 2021; Zhou et al., 2021; Zhang et al., 2021a; Liu et al., 2021; 2022b). However, most methods still apply a similar fusion computation to all inputs, forcing a single policy to compromise among heterogeneous RGB-D reliability relations (Figure 1(b)). Quality-aware approaches partially address this limitation, but a depth-centered score cannot capture the reliability of RGB itself or whether the two modalities provide consistent foreground evidence. Modality Profile (Wang et al., 2023) explicitly describes this

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.

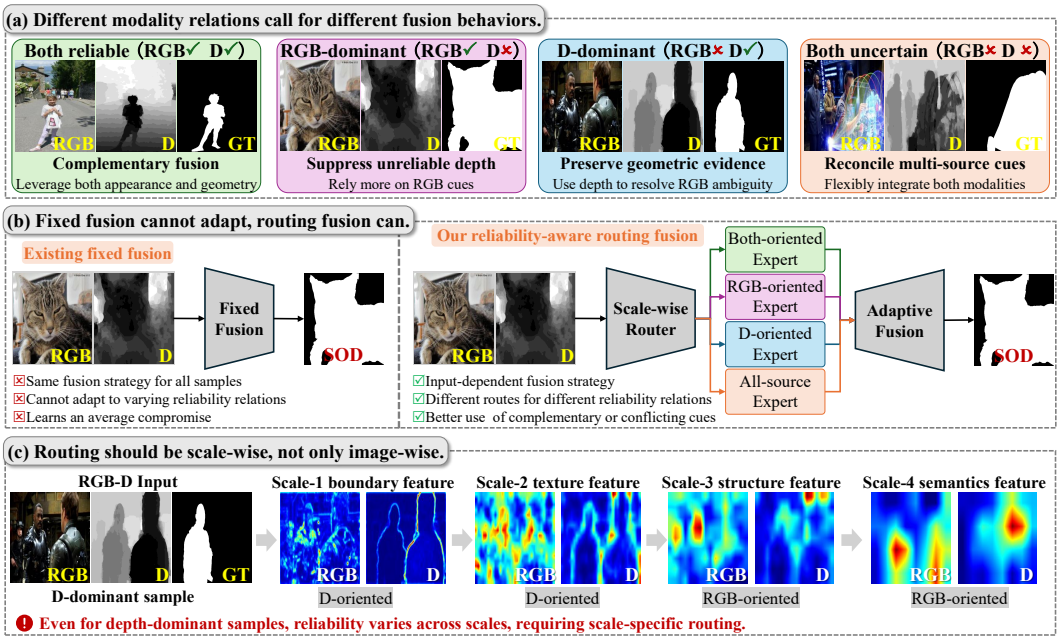


Figure 1: Motivation for scale-wise reliability routing. (a) Different RGB-D reliability relations call for different fusion behaviors: complementary fusion when both modalities are informative, suppression of misleading depth when RGB is more reliable, preservation of geometric evidence when depth resolves RGB ambiguity, and flexible reconciliation when both are uncertain. (b) A fixed fusion rule applies the same computation to all inputs, whereas reliability-aware routing can select and combine specialized fusion paths according to the current modality relation. (c) Reliability is not only sample-dependent but also scale-dependent: even for a globally depth-dominant sample, shallow and deep representations may favor different routes. These observations motivate latent, scale-wise routing rather than a single fixed or image-level fusion policy.

joint relation, yet remains primarily a dataset-level characterization rather than an input-conditioned fusion mechanism.

Image-level adaptation remains insufficient because the relative utility of RGB and depth can vary across feature scales (Figure 1(c)). Shallow features focus on local details such as contours and contrast, where coarse or misaligned depth may introduce boundary errors, whereas deeper features capture semantic structures and can benefit from depth for object-level localization. Therefore, even a globally depth-dominant sample may require RGB-oriented fusion at some scales and depth-oriented fusion at others. Assigning a single image-level reliability state would inevitably overlook these scale-specific differences.

Moreover, standard RGB-D SOD datasets provide no explicit reliability annotations, and reliability is inherently task dependent rather than an intrinsic property of the depth source. A visually smooth depth map may still provide little useful information for the salient target, while an imperfect depth map can remain valuable if its geometry resolves RGB ambiguity. We therefore model modality reliability as a *latent, scale-wise routing variable*: the four reliability relations provide structural priors, while their mixture is learned end-to-end through the saliency and edge objectives.

Based on this formulation, we propose **MPLoRA-MoE**. A shared hierarchical encoder captures common scene semantics, while modality-specific low-rank residuals retain compact appearance- and geometry-dependent transformations. At each scale, a lightweight router softly combines four heterogeneous experts whose input structures encode complementary reliability priors. The routed hierarchy is then decoded by a single edge-semantic pathway. The router predicts neither quality labels nor output masks; it only controls how RGB-D evidence is fused before a common prediction head, making the formulation applicable to generic hierarchical backbones.

Our contributions are threefold:

- We formulate RGB-D fusion as *scale-wise latent reliability routing*, where fusion decisions are learned from task supervision and adapt to both the input sample and feature scale, instead of relying on a fixed fusion rule or a global depth-quality score.
- We instantiate this formulation with modality-specific low-rank adaptation and four heterogeneous experts encoding both-reliable, RGB-oriented, depth-oriented, and uncertainty-aware fusion behaviors, combined by soft routing rather than discrete profile classification.
- We evaluate one model on eight RGB-D SOD benchmarks and use controlled ablations to isolate the effects of structural guidance, expert heterogeneity, routing temperature, and modality-specific adaptation capacity.

2 Related Work

RGB-D saliency and reliability-aware fusion. Early RGB-D SOD established benchmark protocols and handcrafted depth cues (Niu et al., 2012; Peng et al., 2014; Ju et al., 2015). Deep models then shifted toward learning modality-specific representations together with progressively richer interaction. DMRA introduces depth-induced recurrent attention (Piao et al., 2019); BBS-Net separates modality-specific encoding from cross-level aggregation (Fan et al., 2020); JL-DCF jointly learns independent and cooperative representations (Fu et al., 2020); and DCF calibrates multimodal features using depth-related supervision (Ji et al., 2021). Later methods preserve modality specificity (Zhou et al., 2021), suppress redundant cross-modal information (Zhang et al., 2021a), or model long-range interaction with Transformers (Liu et al., 2021; 2022b; Kupyn et al., 2026). These advances primarily improve *how* RGB and depth interact, but usually keep the interaction rule itself shared across samples.

A complementary line addresses unreliable depth through depth-sensitive attention or quality-aware feature manipulation (Sun et al., 2021; Zhang et al., 2021b). Such approaches recognize that depth should not always be trusted, but often summarize reliability from the depth side alone. Modality Profile instead characterizes RGB-D data by the joint quality of both modalities (Wang et al., 2023), which better reflects whether one modality can compensate for the other. We adopt this joint perspective but change its computational role: reliability is treated as a latent, scale-specific variable that directly controls fusion, rather than a dataset label or an explicitly supervised quality score.

Conditional computation and modality-specific adaptation. Mixture-of-experts (MoE) models enable input-dependent computation through learned routing (Shazeer et al., 2017; Fedus et al., 2022). Conventional MoE designs often use structurally interchangeable experts and rely on optimization to induce specialization. In contrast, our experts receive different combinations of raw and cross-modally aligned RGB-D features, so each route is biased toward a distinct fusion behavior while remaining softly composable. Low-rank adaptation (LoRA) captures compact deviations from shared transformations (Hu et al., 2022). This decomposition is complementary to routing: LoRA preserves modality-specific statistics before fusion, whereas MoE determines how those modality representations should interact for the current sample and feature scale.

3 Method

3.1 Overview

Given RGB $I^r \in \mathbb{R}^{3\times H\times W}$ and depth $I^d \in \mathbb{R}^{1\times H\times W}$, MPLoRA-MoE predicts $S \in [0,1]^{H\times W}$. Figure 2 summarizes three stages: shared semantics with modality-specific adaptation, scale-wise latent reliability routing, and a single prediction pathway for all routed features. Thus conditional computation changes the fused representation rather than selecting among multiple output heads.

3.2 Shared encoder with modality-specific LoRA

Independent encoders preserve modality differences but duplicate parameters, whereas complete sharing forces appearance and geometry through identical transformations. We therefore use a shared hierarchical backbone with separate RGB and depth LoRA residuals in the attention projections (q, kv, and output) and MLP layers. Our implementation uses PVT-v2-B5 (Wang et al.,

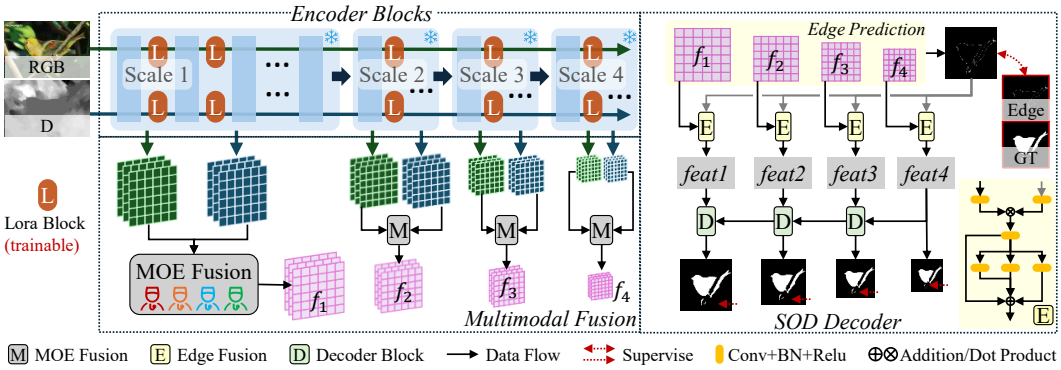


Figure 2: Overview of MPLoRA-MoE. RGB and depth share hierarchical semantics while modality-specific LoRA residuals retain compact appearance- and geometry-dependent transformations. At each scale, a latent router mixes heterogeneous fusion experts before the routed hierarchy is passed to one edge-semantic decoder. Thus, conditional computation changes the fused representation rather than selecting among multiple output heads.

2022), although the routing formulation only requires hierarchical modality features. For a shared projection $W \in \mathbb{R}^{d_o \times d_i}$ and modality $m \in \{r, d\}$,

$$y_m = W x_m + \frac{\alpha}{r} B_m A_m x_m, \tag{1}$$

where $A_m \in \mathbb{R}^{r \times d_i}$ and $B_m \in \mathbb{R}^{d_o \times r}$ are modality-specific, r is the rank, and α scales the residual. The shared term captures transferable scene semantics, while (A_r, B_r) and (A_d, B_d) model compact appearance- and geometry-dependent deviations. The backbone returns four levels with channels $\{64, 128, 320, 512\}$ and strides $\{4, 8, 16, 32\}$. Separate ASPP projections (Chen et al., 2017) map both modalities to a common width $C = 128$, yielding $R_l, D_l \in \mathbb{R}^{C \times H_l \times W_l}$.

3.3 Scale-wise latent reliability routing

The utility of depth can vary across both samples and representation levels, so we instantiate an independent router at each encoder scale l. This design permits shallow stages to reject unreliable depth boundaries while deeper stages still exploit useful geometric structure for object localization. Modality-specific coordinate attention (Hou et al., 2021) first produces $\bar{R}_l = \mathcal{C}_l^r(R_l)$ and $\bar{D}_l = \mathcal{C}_l^d(D_l)$. Bidirectional cross-modal attention then constructs aligned alternatives,

$$\widetilde{R}_l = \mathcal{A}_l^{d \to r}(\bar{R}_l, \bar{D}_l), \qquad \widetilde{D}_l = \mathcal{A}_l^{r \to d}(\bar{D}_l, \bar{R}_l). \tag{2}$$

The router operates on the *unaligned* features $[\bar{R}_l, \bar{D}_l]$. This separation lets the gate estimate the pre-fusion modality relation, while the experts can still choose between raw and cross-modally corrected evidence. Global average pooling and a two-layer MLP produce logits $g_l \in \mathbb{R}^4$, followed by

$$q_{l,k} = \frac{\exp(g_{l,k}/\tau)}{\sum_{j=1}^{4} \exp(g_{l,j}/\tau)}, \qquad k \in \{1, 2, 3, 4\}, \tag{3}$$

where τ controls routing sharpness. No profile or quality target supervises g_l; the routing distribution is learned solely through the downstream saliency and edge objectives. Accordingly, q_l should be interpreted as the task-dependent utility of alternative fusion paths rather than a calibrated estimate of physical sensor quality.

3.4 Reliability-structured experts

Figure 3 shows the four expert paths. The first three use independent 1×1 Conv–BN–ReLU projections, while the fourth uses a 3×3 deformable convolution (Dai et al., 2017). Their inputs are

$$E_l^{(1)} = \mathcal{P}_l^{(1)}([\bar{R}_l, \bar{D}_l]), \qquad E_l^{(2)} = \mathcal{P}_l^{(2)}([\bar{R}_l, \widetilde{D}_l]),$$
$$E_l^{(3)} = \mathcal{P}_l^{(3)}([\widetilde{R}_l, \bar{D}_l]), \qquad E_l^{(4)} = \mathcal{D}_l([\bar{R}_l, \bar{D}_l, \widetilde{R}_l, \widetilde{D}_l]). \tag{4}$$

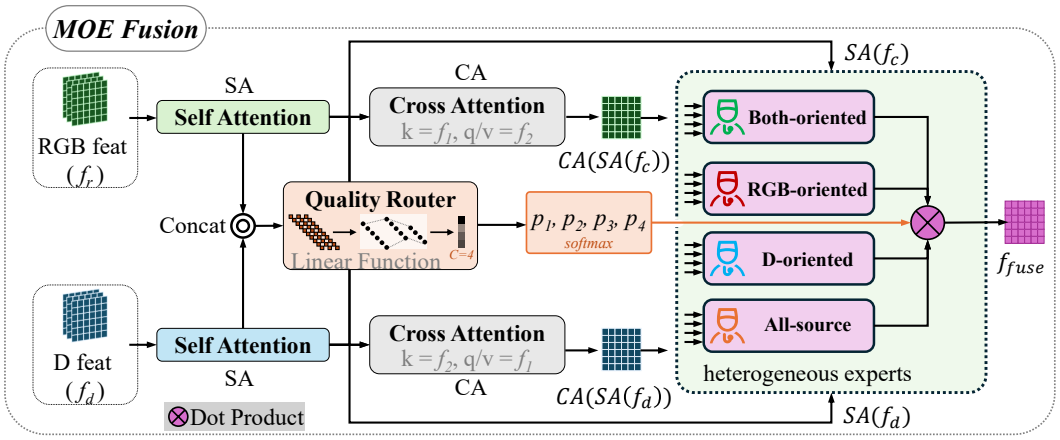


Figure 3: Scale-wise reliability routing (MoE Fusion). Four heterogeneous experts receive different combinations of raw and cross-modally aligned RGB-D features, corresponding to direct, RGB-oriented, depth-oriented, and all-source fusion behaviors. A latent router predicts a soft mixture independently at each feature scale, allowing the same image to use different reliability compositions across representation levels.

The four structures encode complementary reliability priors. Expert 1 preserves direct evidence from both modalities; Expert 2 replaces raw depth with RGB-guided depth; Expert 3 is the symmetric depth-oriented path; and Expert 4 combines all raw and aligned evidence with deformable sampling for stronger disagreement or uncertainty. These semantics are architectural priors rather than supervision targets. The routed feature is

$$F_l = \sum_{k=1}^{4} q_{l,k} E_l^{(k)}, \qquad F_l \in \mathbb{R}^{C \times H_l \times W_l}. \tag{5}$$

Soft mixing lets ambiguous samples use multiple paths, while independent routers allow the image to adopt different reliability compositions across scales. Greater weight on depth-preserving paths corresponds to trusting depth, whereas unreliable depth shifts weight toward the RGB-oriented path.

3.5 Unified Edge-Semantic Decoding and Training

All routed features $\{F_l\}_{l=1}^4$ enter a single decoder adapted from ESCNet (Ye et al., 2025). An edge branch aggregates shallow structural cues with deeper semantics to predict E. Each routed feature is then refined by edge-guided deformable sampling and multi-scale convolutions before coarse-to-fine decoding proceeds from F_4 to F_1; CBAM-style recalibration (Woo et al., 2018) further adjusts channel and spatial responses. Intermediate saliency logits provide deep supervision, while only the final full-resolution prediction is used at inference. This common prediction pathway is important for interpreting the MoE: experts alter the evidence supplied to the decoder, but do not own independent masks or task heads. Hence improvements cannot be attributed to selecting the most favorable output among multiple predictions.

For saliency logit P_i and ground truth G, we use $\mathcal{L}_{\text{str}}(P_i, G) = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{IoU}}$ together with an edge Dice loss $\mathcal{L}_{\text{edge}}(E, G_e)$. The complete objective is

$$\mathcal{L} = \sum_{i=1}^{4} \lambda_i \mathcal{L}_{\text{str}}(P_i, G) + \lambda_e \mathcal{L}_{\text{edge}}(E, G_e), \tag{6}$$

with $\lambda_i = \{0.5, 0.7, 0.9, 1.1\}$ from coarse to fine and $\lambda_e = 1$. No auxiliary reliability-classification loss is used.

Table 1: Quantitative comparison with recent RGB-D SOD methods on eight benchmarks. Numbers in parentheses indicate test-set sizes. M denotes MAE; ↑/↓ indicate higher/lower is better. Best and second-best results are shown in **bold** and underlined, respectively. The table includes both held-out splits of training-source datasets and benchmarks that are entirely unseen during training, allowing comparison under different evaluation regimes.

Method	NJU2K (500)				NLPR (300)				DUT-RGBD (400)				STERE (1,000)			
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$
RD3D	.893	.883	.927	.047	.903	.880	.937	.033	.863	.841	.894	.060	.889	.868	.920	.048
HRTransNet	.917	.920	.952	.032	.931	.925	.966	.019	.918	.925	.951	.033	.915	.912	.954	.032
CAVER	.926	.928	.959	.030	.934	.929	.970	.021	.938	.944	.966	.026	.917	.916	.955	.033
CPNet	.935	**.941**	.964	**.025**	**.940**	**.936**	**.973**	**.016**	.951	**.959**	.975	**.019**	.920	.923	**.960**	**.029**
LAFB	.907	.912	.946	.036	.930	.921	.965	.020	.927	.934	.956	.028	.908	.906	.945	.037
CATNet	.932	.937	.961	.026	**.940**	.934	.972	.018	**.953**	.958	**.976**	**.019**	.921	.922	.958	.030
SATNet	.923	.925	.954	.030	.929	.920	.964	.021	.942	.947	.966	.022	.919	.913	.951	.032
DPPNet	.929	.932	.962	.028	.937	.927	.968	.020	.939	.946	.965	.025	.922	.919	.957	.032
Ours	**.936**	.939	**.965**	.027	**.940**	.932	.971	.018	.947	.956	.973	.021	**.928**	**.925**	.959	**.029**

Method	ReDWeb-S (1,000)				SIP (929)				COME-E (4,600)				COME-H (3,000)			
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$
RD3D	.671	.621	.730	.163	.835	.826	.884	.074	.836	.818	.877	.073	.782	.764	.824	.109
HRTransNet	.724	.710	.800	.127	.860	.876	.916	.056	.858	.856	.909	.056	.816	.818	.869	.084
CAVER	.736	.737	.808	.121	.904	.915	.945	.038	.870	.874	.918	.052	.822	.831	.872	.082
CPNet	.752	.755	.822	.112	.907	.927	.946	**.035**	.884	.889	.928	.045	.834	.854	.889	.071
LAFB	.722	.721	.792	.129	.897	.913	.942	.041	.864	.864	.908	.056	.814	.817	.860	.088
CATNet	.748	.750	.816	.115	.911	.928	.952	**.035**	.882	.897	.932	.043	.834	.855	.890	.071
SATNet	.705	.701	.782	.131	.898	.904	.931	.042	.864	.855	.901	.056	.814	.809	.857	.087
DPPNet	.749	.746	.817	.115	.896	.911	.938	.042	.878	.876	.917	.052	.839	.840	.879	.078
Ours	**.773**	**.784**	**.831**	**.109**	**.915**	**.935**	**.952**	**.035**	**.898**	**.901**	**.935**	**.042**	**.855**	**.863**	**.894**	**.069**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We train on 2,985 labeled RGB-D pairs: 1,485 from NJU2K (Ju et al., 2015), 700 from NLPR (Peng et al., 2014), and 800 from DUT-RGBD (Piao et al., 2019). Evaluation uses NJU2K (500), NLPR (300), DUT-RGBD (400), STERE (1,000) (Niu et al., 2012), ReDWeb-S (1,000) (Liu et al., 2022a), SIP (929) (Fan et al., 2021), COME-E (4,600), and COME-H (3,000) (Zhang et al., 2021a). Following the conventional protocol (Tang et al., 2023), train/test images are disjoint. The first three datasets contribute training samples and use held-out test splits; the remaining five are entirely unseen during training, enabling both within-dataset and cross-dataset evaluation.

Metrics. We report S-measure S_α (Fan et al., 2017), maximum F-measure F_β ($\beta^2 = 0.3$), maximum E-measure E_ϕ (Fan et al., 2018), and MAE; the first three are higher-is-better and MAE is lower-is-better.

Implementation. Experiments run on Ubuntu 22.04 with one NVIDIA RTX 4090 (24 GB). RGB and depth are resized to 416×416 and geometrically augmented jointly; depth is replicated to three channels only at the encoder input. The default model uses PVT-v2-B5, LoRA rank $r = 8$, residual scale $\alpha = 16$, four experts, $\tau = 1$, and decoder width $C = 128$. Training uses AdamW (Loshchilov & Hutter, 2019) for 120 epochs with batch size 4, initial learning rate 5×10^{-5}, weight decay 1.5×10^{-4}, and cosine decay.

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

We compare with RD3D (Chen et al., 2021), HRTransNet (Tang et al., 2023), CAVER (Pang et al., 2023), CPNet (Hu et al., 2024), LAFB (Wang et al., 2024), CATNet (Sun et al., 2024), SATNet (Duan et al., 2025), and DPPNet (Yuan et al., 2025). All use the original dataset depth except SATNet, which substitutes monocular estimates. Table 1 reports all four metrics on eight test sets.

Our largest improvements are observed on cross-dataset evaluations where the test distributions are not directly involved in training. MPLoRA-MoE achieves the best overall perfor-

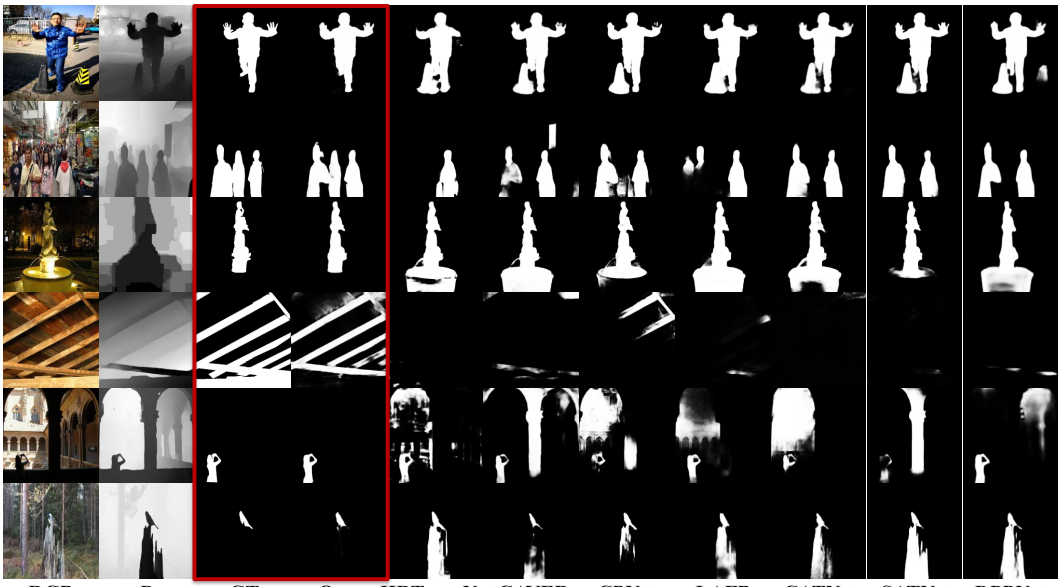


Figure 4: Qualitative comparison under different RGB-D reliability relations. In the first two rows, cluttered RGB appearance makes depth the cleaner cue for foreground localization; in the middle rows, coarse or misleading depth should instead be suppressed in favor of RGB structure; in the final rows, both modalities provide useful evidence. Across these complementary regimes, our method better preserves reliable cues, suppresses modality-specific distractors, and recovers more complete object extent, illustrating the intended input-dependent fusion behavior.

mance across ReDWeb-S, SIP, COME-E, and COME-H, while remaining highly competitive on STERE. ReDWeb-S provides a representative example: our model achieves $S_\alpha/F_\beta/E_\phi/\mathrm{MAE} = .773/.784/.831/.109$, improving the strongest competing values by 2.8%, 3.8%, and 1.1%, respectively, while reducing MAE by 2.7%. A similar trend is observed on COME-H, where our model achieves the best performance across all reported metrics. The consistent improvements in structural similarity, precision–recall quality, alignment quality, and pixel-level error indicate that the gains are not limited to a single evaluation criterion.

The performance gap becomes smaller on held-out splits of datasets that also provide training samples. On NJU2K, our method achieves the best S_α (.936) and E_ϕ (.965), while CPNet remains superior in F_β and MAE; NLPR and DUT-RGBD exhibit similarly close comparisons. This difference suggests that conditional fusion is particularly beneficial when the model encounters more diverse modality reliability relations that cannot be well captured by a single interaction strategy. Since existing benchmarks do not provide explicit reliability annotations, we interpret this cross-dataset behavior as empirical evidence consistent with the proposed reliability-aware routing hypothesis, rather than as a direct measurement of reliability variation.

4.3 Qualitative comparison

Figure 4 covers three representative reliability regimes. In the first two rows, RGB contains substantial clutter, while the relatively clean depth structure provides a more reliable foreground cue. Several competing methods therefore activate background distractors, whereas our model preserves the useful geometric evidence and suppresses these false responses. In the third and fourth rows, depth becomes coarse or misleading; directly propagating such depth information may distort object boundaries, while our predictions retain the more reliable RGB structures. In the final two rows, both modalities provide informative cues, yet some methods still focus on locally high-contrast regions rather than the complete salient object. Our results better preserve object extent and contour integrity while reducing isolated activations. These examples show that the same model can shift among depth-preserving, RGB-oriented, and complementary fusion behaviors, consistent with routing according to relative modality utility rather than applying a single interaction rule to all samples.

4.4 ABLATION STUDIES

Ablations examine reliability-conditioned fusion beyond structural guidance, expert structure versus count, and sensitivity to routing sharpness and modality-specific capacity.

Edge guidance vs. reliability-conditioned fusion. Table 2 isolates the two additions to the shared-encoder baseline. Edge guidance raises F_β from .921 to .924 while MAE stays at .038. Adding MoE increases F_β to .935 and lowers MAE to .035, a further 1.1 percentage-point gain and 7.9% relative MAE reduction. Thus edge supervision refines structure, whereas the larger gain comes from adapting the cross-modal evidence passed to the decoder; the two components are complementary.

Table 2: Component ablation on SIP. M: MAE; ↑/↓: higher/lower is better.

Edge	**MoE**	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$
–	–	.908	.921	.945	.038
✓	–	.910	.924	.948	.038
✓	✓	**.915**	**.935**	**.952**	**.035**

Table 3: Effect of expert number and structure on SIP and COME-H. 4^* uses the proposed heterogeneous experts; the other rows replicate one homogeneous expert architecture. M: MAE; ↑/↓: higher/lower is better. Best/second-best are **bold**/underlined. The superiority of 4^* shows that reliability-oriented specialization matters beyond simply increasing branch count.

K	**SIP**				**COME-H**			
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$
2	.899	.914	.936	.043	.844	.852	.883	.077
4	.909	.927	.950	.037	.852	.858	.892	.071
8	.906	.926	.944	.040	.844	.854	.885	.077
4^*	**.915**	**.935**	**.952**	**.035**	**.855**	**.863**	**.894**	**.069**

Table 4: Sensitivity to routing temperature τ on SIP and COME-H. M: MAE; ↑/↓: higher/lower is better. Best/second-best are **bold**/underlined. The overall optimum around $\tau = 1$ favors a moderate balance between premature hard routing and over-smoothed expert mixing.

τ	**SIP**				**COME-H**			
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$
0.5	.902	.918	.941	.040	.850	.855	.888	.072
0.75	.911	.928	.948	.038	.845	.855	.886	.077
1.0	**.915**	**.935**	**.952**	**.035**	**.855**	**.863**	**.894**	**.069**
1.5	.912	.928	.949	.036	.853	.859	.892	.070
2.0	.912	.924	.947	.036	.852	.859	.892	.072

Expert number and expert structure. Table 3 separates branch count from expert structure. With homogeneous experts, increasing K from 2 to 4 improves both datasets, but $K = 8$ reverses part of the gain (SIP MAE: .043→.037→.040). Replacing homogeneous $K = 4$ with the heterogeneous 4^* configuration improves every metric and lowers MAE to .035 on SIP and .069 on COME-H. The gain therefore comes from distinct direct, RGB-oriented, depth-oriented, and all-source paths rather than simply adding branches.

Routing temperature. Table 4 shows a non-monotonic dependence on τ. On SIP, increasing τ from 0.5 to 1.0 raises F_β from .918 to .935 and lowers MAE from .040 to .035; COME-H follows the same trend. Larger temperatures again weaken several metrics. The optimum near $\tau = 1$ supports soft routing: overly sharp gates encourage premature commitment, whereas overly smooth gates weaken specialization.

LoRA rank. Table 5 shows that modest modality-specific capacity is sufficient. Increasing r from 2 to 8 lowers MAE from .039 to .035 on SIP and from .074 to .069 on COME-H; $r = 16$ gives no consistent gain, while $r = 32$ degrades markedly. This supports the intended decomposition: low-rank residuals capture modality-dependent deviations while most representation capacity remains shared; extra adapter capacity is unnecessary and may hurt generalization.

Table 5: Sensitivity to LoRA rank r on SIP and COME-H. M: MAE; ↑/↓: higher/lower is better. Best/second-best are **bold**/underlined. The results favor moderate rank, indicating that compact modality-specific capacity is sufficient to complement the shared encoder.

r	SIP				COME-H			
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$M \downarrow$
2	.905	.924	.943	.039	.846	.852	.885	.074
4	.910	.926	.946	.037	.851	.856	.890	.071
8	**.915**	**.935**	.952	**.035**	**.855**	**.863**	**.894**	**.069**
16	**.915**	.928	**.953**	**.035**	.854	.861	**.894**	**.069**
32	.908	.927	.946	.038	.839	.846	.880	.078

5 Discussion and Limitations

Interpreting learned reliability. The cross-dataset behavior provides useful evidence for interpreting what the router learns. The gains are generally larger on datasets that are entirely unseen during training, whereas performance is more tightly clustered on held-out splits of training-source datasets. This pattern is consistent with conditional routing becoming more valuable when a fixed fusion rule must accommodate a broader range of RGB-D relations. It also supports a task-dependent view of reliability: rather than estimating depth quality itself, the router learns how useful different fusion paths are for the current prediction.

Accordingly, the learned routing weights should not be interpreted as estimates of physical depth quality. The router receives no quality or profile supervision and is optimized only through the saliency and edge objectives. Its output therefore reflects the *task-relative utility* of alternative fusion paths. A smooth depth map may still be uninformative if it does not distinguish the salient target, whereas an imperfect depth map can remain useful when its coarse geometry resolves ambiguous RGB appearance. Thus, "trusting depth" refers to whether depth provides useful evidence for the task, rather than whether the depth measurement is globally accurate.

This task-relative view of reliability also motivates the scale-wise design. Reliability need not remain constant across the feature hierarchy: shallow features emphasize local contrast and boundary alignment, whereas deeper features capture object-level semantics and coarse localization. The same depth map may therefore be useful at a deeper stage but less reliable near fine boundaries at a shallower stage. Independent soft routers allow these scale-dependent preferences to coexist instead of assigning the entire image a single reliability state. In this sense, learning when to trust depth also means learning *where in the representation hierarchy* depth should influence the prediction.

Limitations. The current formulation has two main limitations. First, expert semantics are encouraged by architectural design rather than explicitly supervised, so different routing paths may remain partially entangled or permute across runs. Second, routing can only reweight and recombine the evidence already present in the two modalities; it cannot recover salient structure that is missing from both RGB and depth. Future work could therefore analyze how scale-wise routing reflects latent modality reliability more directly and improve expert identifiability while preserving the shared prediction pathway.

6 Conclusion

We presented MPLoRA-MoE, which formulates RGB-D fusion as learning a task-dependent, scale-wise latent reliability variable rather than treating depth as uniformly trustworthy. The framework preserves modality-specific information through lightweight adaptation and uses independent routers to combine reliability-structured fusion paths before a shared decoder. Experiments on eight RGB-D SOD benchmarks show strong and consistent performance, while controlled ablations demonstrate that the gains arise from structured expert specialization and adaptive routing rather than simply increasing model capacity. Overall, the results support a simple principle: effective RGB-D fusion depends not only on whether depth is informative, but also on when and at which representation scale it should be trusted.

AI USE STATEMENT

Generative AI tools were used to polish the language of selected manuscript passages and to assist with diagnosing and fixing bugs in the source code. All AI-assisted text was reviewed and edited by the authors, and all AI-suggested code changes were manually inspected and tested before use. The authors take responsibility for the final content of the paper, including text and code produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We provide complete algorithmic and architectural specifications, including loss definitions and hyperparameters, to ensure end-to-end reproducibility. After the review process, we will release the full source code, training/testing scripts, and model checkpoints to enable transparent verification.

REFERENCES

Shilong Bao, Qianqian Xu, Feiran Li, Boyu Han, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. Towards size-invariant salient object detection: A generic evaluation and optimization approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48:1115–1132, 2026.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Qian Chen, Ze Liu, Yi Zhang, Keren Fu, Qijun Zhao, and Hongwei Du. RGB-D salient object detection via 3d convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1063–1071, 2021.

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 764–773, 2017.

Songsong Duan, Xi Yang, Nannan Wang, and Xinbo Gao. Lightweight RGB-D salient object detection from a speed-accuracy tradeoff perspective. *IEEE Transactions on Image Processing*, 34: 2529–2543, 2025.

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.

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 International Joint Conference on Artificial Intelligence*, pp. 698–704, 2018.

Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In *European Conference on Computer Vision*, pp. 275–292, 2020.

Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2075–2089, 2021.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3052–3062, 2020.

Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13713–13722, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

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.

Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, and Li Cheng. Calibrated RGB-D salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9471–9481, 2021.

Ran Ju, Yang Liu, Tongwei Ren, Ling Ge, and Gangshan Wu. Depth-aware salient object detection using anisotropic center-surround difference. *Signal Processing: Image Communication*, 38:115–126, 2015.

Orest Kupyn, Hirokatsu Kataoka, and Christian Rupprecht. S3od: towards generalizable salient object detection with synthetic data. In *International Conference on Learning Representations*, 2026.

Nian Liu, Ni Zhang, Ling Shao, and Junwei Han. Learning selective mutual attention and contrast for RGB-D saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9026–9042, 2022a. doi: 10.1109/TPAMI.2021.3122139.

Zhengyi Liu, Yuan Wang, Zhengzheng Tu, Yun Xiao, and Bin Tang. TriTransNet: RGB-D salient object detection with a triplet transformer embedding network. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4481–4490, 2021.

Zhengyi Liu, Yacheng Tan, Qian He, and Yun 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, 2022b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 454–461, 2012.

Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. CAVER: Cross-modal view-mixed transformer for bi-modal salient object detection. *IEEE Transactions on Image Processing*, 32:892–904, 2023.

Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. RGB-D salient object detection: A benchmark and algorithms. In *European Conference on Computer Vision*, pp. 92–109, 2014.

Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7254–7263, 2019.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Fuming Sun, Peng Ren, Bowen Yin, Fasheng Wang, and Haojie Li. CATNet: A cascaded and aggregated transformer network for RGB-D salient object detection. *IEEE Transactions on Multimedia*, 26:2249–2262, 2024. doi: 10.1109/TMM.2023.3294003.

Peng Sun, Wenhu Zhang, Huanyu Wang, Songyuan Li, and Xi Li. Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1407–1417, 2021.

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.

Kunpeng Wang, Zhengzheng Tu, Chenglong Li, Cheng 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.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVT v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

Xuehao Wang, Shuai Li, Chenglizhao Chen, Aimin Hao, and Hong Qin. Modality profile: A new critical aspect to be considered when generating RGB-D salient object detection training set. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3355–3364, 2023. doi: 10.1145/3581783.3611985.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *European Conference on Computer Vision*, pp. 3–19, 2018.

Sheng Ye, Xin Chen, Yan Zhang, Xianming Lin, and Liujuan Cao. ESCNet: Edge-semantic collaborative network for camouflaged object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20053–20063, 2025.

Junbin Yuan, Yiqi Wang, Zhoutao Wang, Qingzhen Xu, Bharadwaj Veeravalli, and Xulei Yang. DPPNet: A depth pixel-wise potential-aware network for RGB-D salient object detection. *IEEE Transactions on Multimedia*, 27:4256–4268, 2025. doi: 10.1109/TMM.2025.3535386.

Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. RGB-D saliency detection via cascaded mutual information minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4338–4347, 2021a.

Wenbo Zhang, Ge-Peng Ji, Zhuo Wang, Keren Fu, and Qijun Zhao. Depth quality-inspired feature manipulation for efficient RGB-D salient object detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 731–740, 2021b.

Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, and Ling Shao. Specificity-preserving RGB-D saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4681–4691, 2021.