Uncertainty-Aware Modality Fusion for Unaligned RGB-T Salient Object Detection

Supplementary Material

7. Comparison on Light Field Dataset

We conduct a comprehensive comparison between our proposed approach and several state-of-the-art light field salient object detection methods, including DLGLRG, PANet, Auto-MSFNet, PoolNet+, LFTransNet, BBRF, MENet, LF-Tracy, PCNet, and CamoDiffusion. Two widely used light field datasets, DUTLF-V2 and PKU-LF, are employed for quantitative evaluation. Specifically, DUTLF-V2 contains 2,957 training samples and 1,247 testing samples covering ten representative object categories across a wide range of real-world light field scenes. In contrast, PKU-LF is the largest publicly available light field dataset to date, including over one hundred object categories with 3,500 training images and 1,500 testing images, providing a more diverse benchmark for evaluating the generalization ability of light field SOD models.

Table 6 summarizes the quantitative results in terms of three commonly used metrics: the Structure-measure (S_m), the weighted F-measure (F_β^ω), and the Enhanced-alignment measure (E_m). As shown, our method consistently achieves the best performance across all metrics on both datasets, demonstrating its strong ability to handle complex light field cues, angular variations, and background interference. These results validate the superior capability of our framework in generating spatially consistent and perceptually aligned salient maps, even in challenging light field conditions.

Table 6. Quantitative comparisons on two light field datasets. The best results are marked in bold.

Method	Source	DUTLF-V2			PKULF		
		$S_m \uparrow$	$F_\beta^\omega \uparrow$	$E_m \uparrow$	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$E_m \uparrow$
DLGLRG	ICCV21	0.861	0.778	0.902	0.856	0.763	0.897
PANet	TCyB21	0.863	0.784	0.907	0.865	0.775	0.907
Auto-MSFNet	ACMMM21	0.847	0.786	0.891	0.877	0.823	0.913
PoolNet+	TPAMI22	0.827	0.724	0.869	0.831	0.793	0.867
LFTransNet	TCSVT23	0.896	0.843	0.936	0.897	0.842	0.938
BBRF	TIP23	0.869	0.820	0.905	0.902	0.863	0.932
MENet	CVPR23	0.857	0.792	0.884	0.888	0.907	0.911
MINet	TII24	0.805	0.709	0.877	0.818	0.712	0.875
LF-Tracy	ICPR25	0.868	0.808	0.920	0.899	0.851	0.938
PCNet	AAAI25	0.872	0.835	0.921	0.891	0.850	0.935
CamoDiffusion	TPMI25	0.911	0.873	0.939	0.910	0.868	0.936
UMFNet	-	**0.926**	**0.877**	**0.966**	**0.940**	**0.914**	**0.970**

Visual comparisons on representative examples are illustrated in Fig. 5. It can be observed that our approach generates more complete and visually consistent salient regions with finer structural boundaries, while other competing methods tend to either miss small-scale details or produce background artifacts. Notably, our model demonstrates superior robustness in complex light field scenarios characterized by occlusion, defocus blur, and non-uniform illumination.

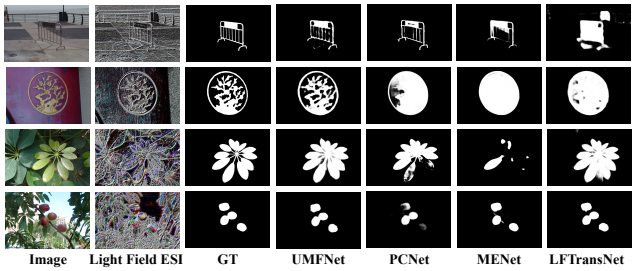

Figure 5. Qualitative comparisons of the state-of-the-art algorithms with our approach on light field datasets.

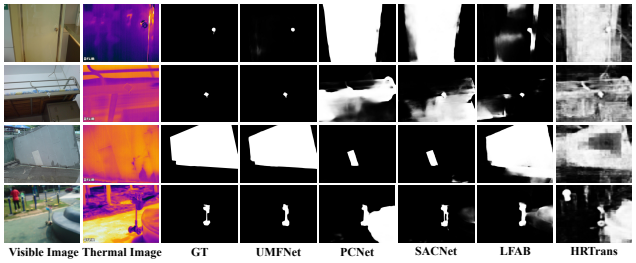

Figure 6. Qualitative comparisons of the state-of-the-art algorithms with our approach on UVT2000 dataset.

8. More Visualizations

8.1. Visualization on UVT2000

To further demonstrate the generalization capability of our model, we provide qualitative results on the UVT2000 RGB-T dataset. As illustrated in Fig. 6, our method produces accurate and consistent saliency maps across diverse illumination conditions and thermal intensities. The results show that our model effectively leverages complementary cues from the visible and thermal modalities, achieving robust detection even under weak illumination or background clutter. This verifies its strong adaptability to varying cross-modal distributions and scene complexities.

8.2. Visualization on Aligned Dataset

To evaluate cross-domain adaptability, we visualize our results on several aligned RGB-T datasets (VT1000) captured under different environmental and sensor conditions. As

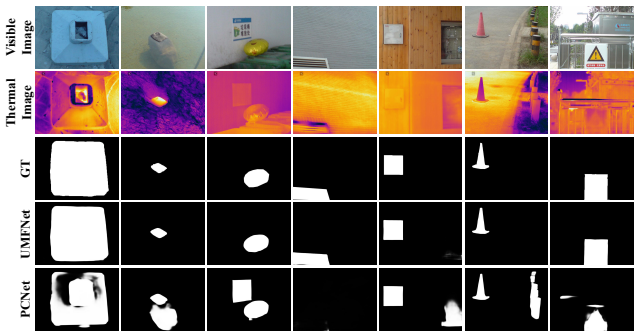

Figure 7. Qualitative comparisons of the state-of-the-art algorithms with our approach on aligned dataset.

shown in Fig. 7, our approach maintains structural consistency and visual quality across scenes with significant thermal–visible modality gaps. The predicted saliency maps remain stable when the thermal signal is weak or when visible information is degraded, highlighting the robustness of our model in handling cross-modality imbalance and heterogeneous background interference.

8.3. Feature Visualization

To gain deeper insights into the internal representations of our RGB-T model, we visualize the intermediate feature responses, as shown in Fig. 8. The first row displays the RGB and thermal inputs, ground truth, and boundary annotations, while the second row illustrates the saliency predictions from shallow to deep layers. It can be observed that the predicted saliency progressively evolves from coarse object localization to fine-grained boundary refinement. Rows three to six further show the corresponding feature activations across different network stages. Shallow layers mainly focus on low-level textures and edges, whereas deeper layers capture high-level semantic and structural information, highlighting complete and coherent object regions. These visualizations clearly demonstrate that our network effectively integrates complementary cues from the visible and thermal modalities, achieving robust and discriminative feature representations under challenging cross-modal conditions.

9. Lightweight Attention Mechanism in the Uncertainty Alignment Module

The lightweight attention mechanism in the Uncertainty Alignment Module (UAM) is implemented through three compact transformer components, namely *transformer_rgb*, *transformer_t*, and *transformer_two*. Each module estimates pixel-wise Gaussian parameters $\mu(p)$ and $\sigma^2(p)$ to jointly capture the deterministic feature content and its uncertainty. Unlike conventional transformer-based fusion schemes that rely on global self-attention, our design adopts

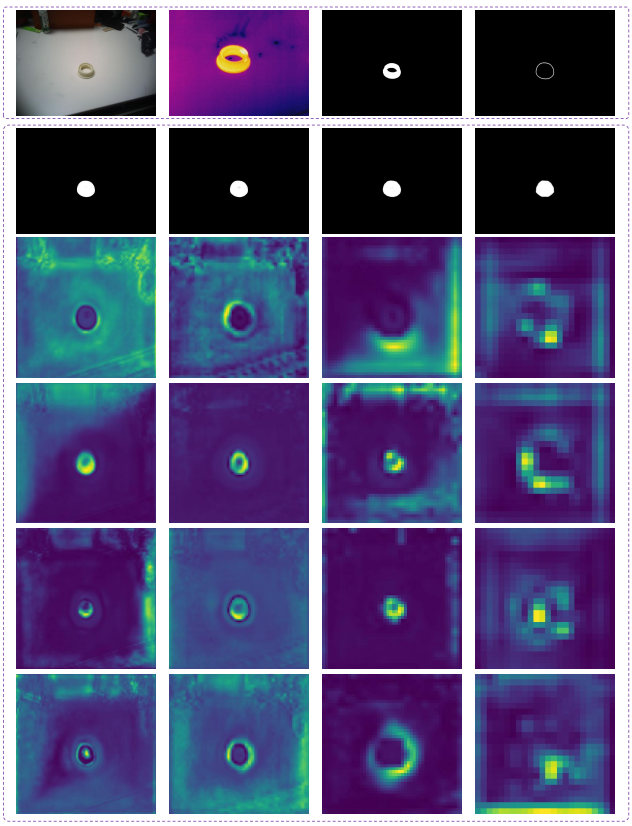

Figure 8. Feature visualization of the proposed network. Row 1: RGB, thermal, ground truth, and boundary maps. Row 2: saliency predictions from shallow to deep layers. Rows 3–6: feature maps at different depths, showing the transition from detailed textures to semantic structures.

localized depthwise filtering and channel-wise MLP refinement, which effectively balance computational efficiency and representational capacity.

The *transformer_rgb* and *transformer_t* modules share an identical architecture for processing the visible and thermal modalities, respectively. Given an input feature map $F_M \in \mathbb{R}^{C\times H\times W}$, a depthwise convolutional attention layer first performs local spatial aggregation followed by normalization:

$$x' = F_M + \mathrm{DWConv}(\mathrm{BN}_1(F_M)), \tag{10}$$

where DWConv denotes a 3×3 depthwise convolution capturing fine-grained spatial dependencies, and BN_1 represents batch normalization for training stability. Two lightweight MLP branches are then used to estimate the Gaussian parameters:

$$\mu_M = x' + \mathrm{MLP}_{\mu}(\mathrm{BN}_2(x')), \tag{11}$$

$$\log \sigma_M^2 = x' + \mathrm{MLP}_{lv}(\mathrm{BN}_3(x')). \tag{12}$$

The variance term is clipped within $[-10, 10]$ for numerical stability, and the standard deviation is obtained as

$$\sigma_M = \exp\big(0.5 \cdot \log \sigma_M^2\big). \tag{13}$$

This yields a pixel-wise Gaussian distribution $\mathcal{N}(\mu_M, \sigma_M^2)$ that encodes both the semantic estimation and local uncertainty within each modality. The reparameterized latent representation is computed via:

$$\tilde{z}_{F_M} = \mu_M + \sigma_M \odot \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I), \tag{14}$$

which allows differentiable sampling of stochastic features during training. To prevent overconfident activations, a mean-based KL divergence regularization term is applied to constrain the learned Gaussian toward a standard normal prior $\mathcal{N}(0, 1)$, stabilizing uncertainty propagation within each modality.

To further achieve structural alignment between modalities, we construct a conditionally fused latent distribution based on the sampled representations $\tilde{z}_{F_v}$ and $\tilde{z}_{F_t}$. The *transformer_two* module models their joint dependencies and infers a cross-modal Gaussian distribution:

$$z_{F_{vt}} \sim \mathcal{N}(\mu_{F_{vt}}, \sigma_{F_{vt}}^2), \tag{15}$$

$$\tilde{z}_{F_{vt}} = \mu_{F_{vt}} + \sigma_{F_{vt}} \odot \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I). \tag{16}$$

In practice, $\tilde{z}_{F_v}$ and $\tilde{z}_{F_t}$ are concatenated along the channel dimension and processed through a depthwise cross-attention block:

$$x = \tilde{z}_{F_v} + \mathrm{DWConv}(\mathrm{GELU}(\mathrm{BN}(\mathrm{Conv}_{1\times1}([\tilde{z}_{F_v}, \tilde{z}_{F_t}])))). \tag{17}$$

This residual design enables the visible modality to selectively integrate complementary structural cues from the thermal modality. Two MLP branches then produce the cross-modal Gaussian parameters:

$$\mu_{F_{vt}} = x + \mathrm{MLP}_{\mu}(\mathrm{BN}_2(x)), \tag{18}$$

$$\log \sigma_{F_{vt}}^2 = x + \mathrm{MLP}_{lv}(\mathrm{BN}_3(x)). \tag{19}$$

Here, $\mu_{F_{vt}}$ encodes the semantically aligned representation across modalities, while $\sigma_{F_{vt}}^2$ quantifies the residual uncertainty due to modality discrepancy. Unlike the unimodal latent features $\tilde{z}_{F_v}$ and $\tilde{z}_{F_t}$ that primarily preserve intra-modal structure, the fused latent variable $\tilde{z}_{F_{vt}}$ captures semantic coherence and shared spatial structures in the latent Gaussian space.

Finally, the *transformer_fusion* module integrates $\tilde{z}_{F_t}$ and $\tilde{z}_{F_{vt}}$ under an uncertainty-guided fusion strategy to produce the aligned infrared feature map $\tilde{F}_t$. The fusion process is formulated as:

$$x_f = \tilde{z}_{F_t} + \mathrm{DWConv}(\mathrm{BN}(\mathrm{Conv}_{1\times1}([\tilde{z}_{F_t}, \tilde{z}_{F_{vt}}]))), \tag{20}$$

Table 7. Pseudocode of the forward process in the Uncertainty Alignment Module (UAM). During training, latent representations are sampled using the reparameterization trick, while in inference, deterministic means are used for stable prediction.

Step	Operation
1	Input visible feature F_v and thermal feature F_t extracted from the backbone.
2	Estimate modality-wise Gaussian parameters: $(\mu_{F_v}, \log \sigma_{F_v}^2, \sigma_{F_v}) \leftarrow$ transformer_rgb(F_v) $(\mu_{F_t}, \log \sigma_{F_t}^2, \sigma_{F_t}) \leftarrow$ transformer_t(F_t)
3	**If training:** $\tilde{z}_{F_v} = \mu_{F_v} + \sigma_{F_v} \odot \mathcal{N}(0, I), \quad \tilde{z}_{F_t} = \mu_{F_t} + \sigma_{F_t} \odot \mathcal{N}(0, I)$ **Else:** $\tilde{z}_{F_v} = \mu_{F_v}, \quad \tilde{z}_{F_t} = \mu_{F_t}$
4	Construct cross-modal latent distribution: $(\mu_{F_{vt}}, \log \sigma_{F_{vt}}^2, \sigma_{F_{vt}}) \leftarrow$ transformer_two$(\tilde{z}_{F_v}, \tilde{z}_{F_t})$ **If training:** $\tilde{z}_{F_{vt}} = \mu_{F_{vt}} + \sigma_{F_{vt}} \odot \mathcal{N}(0, I)$ **Else:** $\tilde{z}_{F_{vt}} = \mu_{F_{vt}}$
5	Generate the aligned feature map through uncertainty-guided fusion: $\tilde{F}_t =$ transformer_fusion$(\tilde{z}_{F_t}, \tilde{z}_{F_{vt}})$
Output:	Aligned feature $\tilde{F}_t$

$$x_f = x_f + \mathrm{MLP}(\mathrm{BN}(x_f)). \tag{21}$$

This step adaptively aggregates deterministic cues from μ_{F_t} and $\mu_{F_{vt}}$ while modulating their relative contributions according to the uncertainty maps $\sigma_{F_t}^2$ and $\sigma_{F_{vt}}^2$. The resulting representation x_f is mapped to $\tilde{F}_t$, completing the latent-space alignment process between the two modalities.

Overall, the three transformer components (*transformer_rgb*, *transformer_t*, and *transformer_two*) collectively implement a lightweight attention mechanism that replaces heavy global self-attention with localized depth-wise filtering and compact channel-wise refinement. As summarized in Table 7, the forward process alternates between modality-specific Gaussian estimation, reparameterized sampling, and cross-modal fusion. By explicitly modeling pixel-wise Gaussian distributions (μ_M, σ_M^2) and propagating uncertainty through the fusion hierarchy, the UAM achieves stable, interpretable, and efficient cross-modal alignment, enhancing robustness under complex multimodal conditions.