Breaking Alignment Barriers: TPS-Driven Semantic Correlation Learning for Alignment-Free RGB-T Salient Object Detection

Lupiao Hu[1], Fasheng Wang[1]*, Fangmei Chen[1], Fuming Sun[1], Haojie Li[2]

[1]Dalian Minzu University, Dalian China
[2]Shandong University of Science and Technology, Qingdao China
1527155801@qq.com, {wangfasheng,20161358,sunfuming}@dlnu.edu.cn, hjli@sdust.edu.cn

Abstract

Existing RGB-T salient object detection methods predominantly rely on manually aligned and annotated datasets, struggling to handle real-world scenarios with raw, unaligned RGB-T image pairs. In practical applications, due to significant cross-modal disparities such as spatial misalignment, scale variations, and viewpoint shifts, the performance of current methods drastically deteriorates on unaligned datasets. To address this issue, we propose an efficient RGB-T SOD method for real-world unaligned image pairs, termed Thin-Plate Spline-driven Semantic Correlation Learning Network (TPS-SCL). We employ a dual-stream MobileViT as the encoder, combined with efficient Mamba scanning mechanisms, to effectively model correlations between the two modalities while maintaining low parameter counts and computational overhead. To suppress interference from redundant background information during alignment, we design a Semantic Correlation Constraint Module (SCCM) to hierarchically constrain salient features. Furthermore, we introduce a Thin-Plate Spline Alignment Module (TPSAM) to mitigate spatial discrepancies between modalities. Additionally, a Cross-Modal Correlation Module (CMCM) is incorporated to fully explore and integrate inter-modal dependencies, enhancing detection performance. Extensive experiments on various datasets demonstrate that TPS-SCL attains state-of-the-art (SOTA) performance among existing lightweight SOD methods and outperforms mainstream RGB-T SOD approaches.

Code — https://github.com/HTUTU2/TPS-SCL

Introduction

Existing RGB-T salient object detection (SOD) methods rely on aligned RGB-T datasets, which are typically generated through manual alignment processes. This situation forms an obstacle to the development of real-world applications, as RGB-T image pairs captured directly by devices are often unaligned. In such raw image pairs, salient objects tend to exhibit substantial spatial and scale misalignment, resulting in weak correlations between the RGB and thermal modalities. Consequently, these weak correlations hinder the effective extraction of complementary information and guidance cues for SOD.

*Corresponding author.

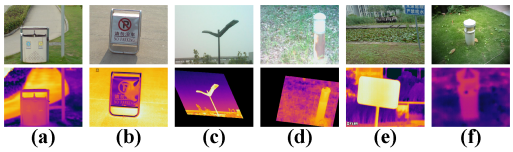

Figure 1: Samples of aligned (a)(b), weakly aligned (c)(d), and unaligned (e)(f) image pairs.

Tu et al. (Tu et al. 2022) are the first to focus on addressing the issue of weakly aligned image pairs. They applied random affine transformation to existing aligned datasets to artificially generate weakly aligned datasets (as shown in Fig. 1, (c) and (d)) and proposed a DCNet for weakly aligned RGB-T SOD. Although DCNet achieved promising detection performance, the local relationships built by its dynamic convolutions are insufficient to handle the large spatial misalignments commonly found in real-world scenarios (as shown in Fig. 1 (e), (f)). To tackle this problem, Wang et al. (Wang et al. 2024b) released the first unaligned RGB-T dataset, UVT2000, and proposed a correlation modeling method based on asymmetric windows. Subsequently, Wang et al.(Wang et al. 2025a) constructed UVT20K, the largest unaligned RGB-T dataset, featuring multiple challenging attributes. They introduced homography estimation to reduce cross-modal discrepancies and employed an attention mechanism to propagate inter-modal correlations throughout each modality. However, homography estimation cannot handle local deformations or nonlinear variations within images. It is ineffective in addressing the spatial misalignment and local deformation caused by significant viewpoint changes during the raw image capturing process by sensors.

In this work, we introduce a Thin Plate Spline (TPS (Duchon 1977)) driven Semantic Correlation Learning for Alignment-Free RGB-T SOD (TPS-SCL), which is capable of handling complex nonlinear deformations and effectively resolving spatial misalignments and image distortions in unaligned image pairs. We design a TPS alignment module that warps and maps the salient objects in the thermal modality to the RGB coordinate space, thereby explicitly aligning the common regions between the RGB and thermal modalities. However, due to the severe variations in spatial location, scale, and viewpoint between the unaligned dual-

modal inputs, directly aligning the raw features proves to be suboptimal. As the features are progressively downsampled, the semantic differences in high-level features are significantly reduced. Therefore, we propose a Semantic Correlation Constraint Module (SCCM) that performs preliminary correlation modeling on the highest-level semantic features to guide and constrain low-level features to focus on the globally salient objects while suppressing redundant background noise. In addition, we design a Cross-Modal Correlation Module (CMCM) to fully explore and exploit the correlations between salient regions across the two modalities. Specifically, we project the features from both modalities into a shared hidden state space and employ a gated mechanism to perform dual hidden state transformations for cross-modal deep feature fusion. This approach further reduces modality discrepancies and enhances the accuracy of saliency prediction.

In addition, due to the linear complexity and low parameter count of Efficient VMamba (Wang et al. 2025b), we build our correlation modeling method based on it. Leveraging its strong long-sequence modeling capability, we capture contextual dependencies across modalities and effectively explore and integrate cross-modal cues. This design strikes a balance between parameter count, computational complexity, and detection accuracy, enabling the model to achieve reliable detection performance on raw alignment-free RGB-T image pairs with lower computational complexity and parameter count.

Our contribution can be summarized as follows:

- We propose a TPS-driven Semantic Correlation Learning network, which is designed to handle unaligned RGB-T image pairs in real-world scenarios by deeply mining saliency cues for accurate detection.
- We introduce a SCCM that utilizes high-level semantic information to constrain hierarchical features, effectively enhancing attention to global salient objects and suppressing background noise.
- We propose a TPSAM, which enhances local structural perception through Local Mamba's localized window scanning and integrates TPS transformation to precisely align co-salient regions across modalities, significantly reducing the impact of spatial discrepancies.
- We design a CMCM that captures inter-modal correlations via an interactive gated mechanism for hidden state transformation across modalities, thereby improving saliency prediction accuracy.

Related Works

RGB-T SOD for Aligned Data

The emergence of RGB-T datasets (VT821, VT1000, and VT5000) has greatly promoted the prosperity of RGB-T SOD. Cong et al. (Cong et al. 2023) designed a Global Illumination Estimation Module to re-evaluate the role of the thermal modality in the SOD task and enrich the semantic information of thermal images, making them more suitable for saliency detection. Building on this, Song et al. (Song et al. 2024) further considered the impact of illumination conditions and proposed a Salient Illumination-Aware Estimator to assess the intensity and distribution of illumination within RGB-T image pairs. Wang et al. (Wang et al. 2024a) proposed a Weight Generation Module to compute the unique contribution weights of the two modalities and guide the following complementary fusion. Wang et al. (Wang et al. 2024c) designed an Adaptive Fusion Repository, which is embedded into the network hierarchy to fully integrate the complementary information from different modalities. Tang et al. (Tang et al. 2025) proposed a Divide-and-Conquer Strategy-based Triple-Stream Network, which employs three separate streams to explore and integrate cues from RGB and thermal modalities.

RGB-T SOD for unaligned Data

Recently, Tu et al. (Tu et al. 2022) proposed a weakly aligned dataset and addressed the weak correlation problem in weakly aligned image pairs through affine transformation and dynamic convolution. While the approach achieved promising results, affine transformation and dynamic convolution are insufficient to effectively handle large spatial deviations caused by significant viewpoint changes. To address this issue, Wang et al. (Wang et al. 2024b) designed a pair of asymmetric windows to model the correlation information in unaligned image pairs. They further incorporated deformable convolutions in the decoding process to reduce spatial discrepancies between the unaligned modalities. However, the fixed asymmetric windows lack the flexibility to adaptively model cross-modal correlations in diverse scenes, and they tend to introduce a significant amount of irrelevant noise. Building on this, Wang et al. (Wang et al. 2025a) designed a semantics-guided homography estimation module, which estimates a homography matrix to align RGB and thermal features, thereby reducing cross-modal discrepancies and facilitating subsequent correlation modeling. However, the homography matrix is insufficient for handling complex local deformations and requires the integration of a semantic adapter to adapt to RGB-T datasets, which inevitably increases computational overhead.

Proposed Method

As shown in Fig. 2, our TPS-SCL takes a pair of unaligned RGB-T images as input and consists of two parallel encoders, a SCCM module, a TPSAM module, a CMCM module, and a decoder. Specifically, the two encoders (MobileViT-S (Mehta and Rastegari 2022)) respectively extract multi-level features from the input image pair (denoted as I_{rgb} and I_t), which are represented as F_m^i, $m \in \{rgb, t\}$, $i = 1, 2, 3, 4$. The SCCM leverages an efficient scanning mechanism (Wang et al. 2025b) to perform initial correlation modeling on high-level semantic features, which in turn guides the shallow layers to focus on salient information and enhances attention to globally salient objects. Due to spatial misalignment, scale variation, and viewpoint rotation between corresponding objects in unaligned image pairs, the TPSAM adaptively aligns salient regions from T to the RGB modality using dynamic control points, thereby

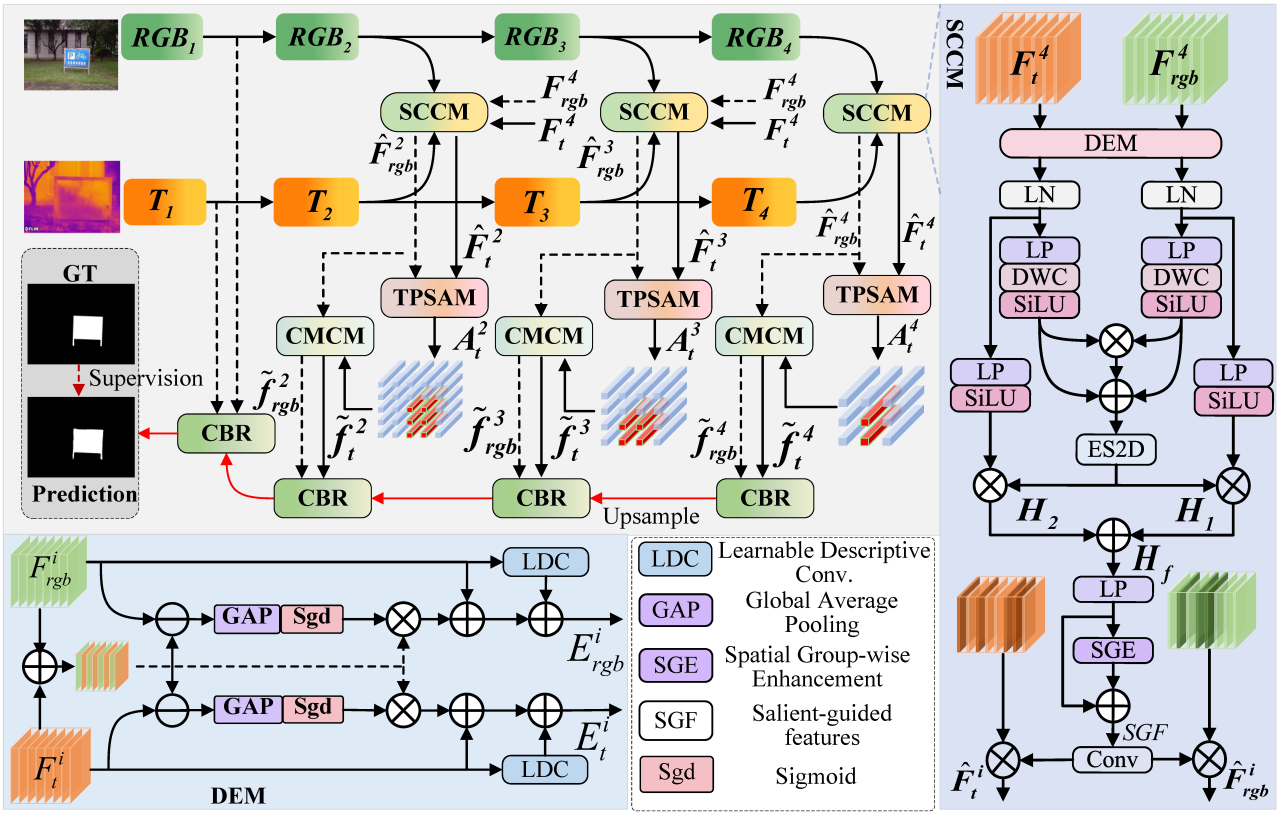

Figure 2: Overall structure of the proposed TPS-SCL.

reducing cross-modal discrepancies. Building on this, the CMCM models the inter-modal correlation between the two modalities, effectively mining and utilizing their correlations and complementary cues to improve detection accuracy. Finally, the decoder integrates the multi-level features output by CMCM to generate the final saliency prediction.

Semantic Correlation Constraint Module

In unaligned RGB-T image pairs, co-salient objects often exhibit discrepancies in spatial location and scale. Directly aligning bimodal features introduces interference from non-significant information, further amplifying spatial differences and hindering the model's learning of a unified cross-modal representation. To address this, we model the correlation between the high-level semantic features of the two modalities and generate saliency-guided features (SGF) to constrain cross-modal related information within salient regions and suppress background noise.

As shown on the right of Fig. 2, SCCM takes the top-level features (F^4_{rgb} and F^4_t) as input. To compensate for the potential loss of local information caused by ES2D, the input top-level features are processed through a differential enhancement module (DEM, bottom left of Fig. 2): $E^4_{rgb/t} = DEM(F^4_{rgb/t})$.

Subsequently, the differentially enhanced features, which retain modality-specific and complementary information, are further explored to model the correlation between different modalities. $E^4_{rgb/t}$ are first fused to generate a shared feature representation H:

$$\begin{aligned} H_{rgb} &= SiLU(DWC(LP(LN(E^4_{rgb})))) \\ H_t &= SiLU(DWC(LP(LN(E^4_t)))) \\ H &= H_{rgb} \oplus H_t \oplus H_{rgb} \otimes H_t \end{aligned} \tag{1}$$

where $DWC(.)$ denotes a depthwise convolution operation, LP is the linear projection, LN refers to layer normalization, and $\otimes$ indicates element-wise multiplication. The shared feature H is then passed through an ES2D layer to capture long-range spatial dependencies across the two modalities. To suppress redundant channel information introduced by the multiple hidden layers of ES2D, the output features are further refined using a residual connection with a lightweight Spatial Group-wise Enhancement (SGE (Li, Hu, and Yang 2019)) attention mechanism.

$$\begin{aligned} H_1 &= ES2D(H) \otimes SiLU(LP(LN(E^4_{rgb}))) \\ H_2 &= ES2D(H) \otimes SiLU(LP(LN(E^4_t))) \\ SGF &= SGE(LP(H_1 \oplus H_2)) \oplus LP(H_1 \oplus H_2) \end{aligned} \tag{2}$$

The output saliency-guided map is upsampled using a 3×3 convolution to match the channel dimensions and spatial resolution of features at different layers. It is then element-wise multiplied with each corresponding layer's features to constrain shallow salient semantic information to focus on co-salient regions and suppresses background noise interference. This can be formulated as follows:

$$\hat{F}^i_{rgb} = UP_{2^{4-i}}(Conv(SGF)) \otimes F^i_{rgb}$$
$$\hat{F}^i_t = UP_{2^{4-i}}(Conv(SGF)) \otimes F^i_t \quad (3)$$

where $UP_k(.)$ denotes k-times upsampling via bilinear interpolation, and $i = 2, ..., 4$.

The detailed process of the DEM is straightforward (Fig. 2 bottom left).

TPS Alignment Module

Although the two modalities undergo high-level semantic constraint and mutual enhancement through the SCCM module, they remain misaligned. Directly modeling the correlation between these unaligned features may lead to mismatches and inaccurate identification of salient regions. To address this issue, we propose the TPSAM module. As illustrated in Fig. 3, it employs TPS to warp the salient regions of the thermal modality into the RGB spatial coordinate system, thereby reducing spatial discrepancies.

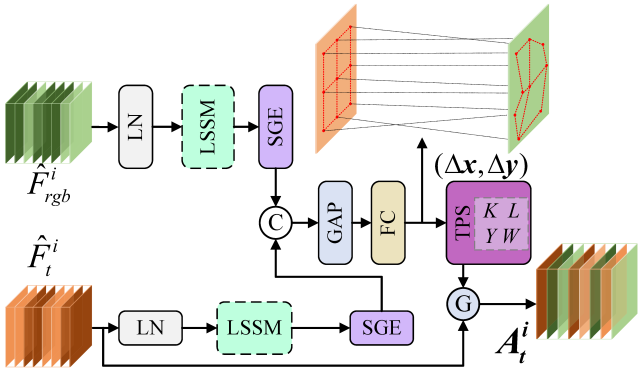

Figure 3: Structure of TPSAM.

The SCCM primarily relies on high-level semantic features, which may overlook spatial contextual information from lower-level features, resulting in incomplete feature representations. To address this, TPSAM first takes the enhanced RGB and thermal features as input and applies Local Scanning State Machine (LSSM) (Huang et al. 2025) with local window scanning to capture global context across windows, and enhance local object details. Afterward, a lightweight Spatial Group-wise Enhancement (SGE) attention mechanism is applied to suppress redundant information. This process can be expressed as follows:

$$\tilde{E}^i_{rgb} = SGE(LSSM(LN(\hat{F}^i_{rgb})))$$
$$\tilde{E}^i_t = SGE(LSSM(LN(\hat{F}^i_t))) \quad (4)$$

The enhanced features are concatenated and passed through GAP to obtain global features which are then fed into a FC layer to predict the displacement of each control point in the source image, thereby dynamically updating the x and y coordinates of the control points in the target image.

$$(\Delta x, \Delta y) = FC(GAP(Concat(\tilde{E}^i_{rgb}, \tilde{E}^i_t)))$$
$$Q(x_2, y_2) = P(x + \Delta x, y + \Delta y) \quad (5)$$

where $(\Delta x, \Delta y)$ represents the displacement of source control points along the x and y axes, and Q is the coordinate matrix of the target control points. The core idea of TPS is to achieve a smooth spatial mapping by minimizing the bending energy. The process begins by constructing an initial point grid P uniformly sampled in the interval [-1,1]. By adding the predicted displacements to this grid, we obtain the target control point coordinate matrix Q. Subsequently, the transformation parameters are computed based on these target control points. The first step is to construct the distance matrix K:

$$K_{ij} = \|p_i - p_j\|^2 \log(\|p_i - p_j\|^2) \quad (6)$$

where p_i represents the coordinates of the i-th source control point in the control point matrix. Next, the augmented matrix L is constructed as follows:

$$L = \begin{bmatrix} K & P_{aug} \\ P^\top_{aug} & 0 \end{bmatrix} \quad (7)$$

where $P_{aug} = [1, P]$ denotes the augmented source control point matrix. Finally, the target matrix Y is constructed as $Y = [Q\ 0]^\top$, where Q is the target control point matrix. The transformation parameters W, which include both the radius basis function (RBF) weights and affine coefficients, are solved using a pseudoinverse $W = L\dagger Y$, where $\dagger$ denotes the pseudoinverse operation. Subsequently, the transformation network G is generated based on the transformation parameters, resulting in a smooth mapping function that minimizes bending energy. This function transforms the coordinates of salient regions from the thermal modality to the corresponding salient regions in the RGB modality. The process is formulated as:

$$R = \sum_{i-1}^{N} w_i U(\|X - p_i\|) + \rho_0 + \rho_1 x + \rho_2 y \quad (8)$$

where R denotes the transformed target point coordinates, ρ_i denotes affine coefficient, X denotes any point in the source image, and x and y refer to the x-axis and y-axis coordinates of X, respectively. $U(r) = r^2 \log(r)$ is the RBF that controls the smoothness of the transformation. Finally, the TPS transformation is applied to obtain the warped thermal image $A^i_t = G(\hat{F}^i_t)$.

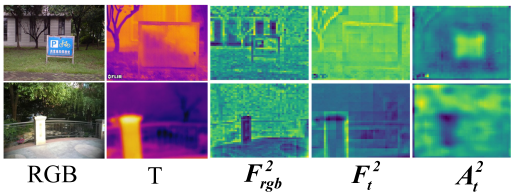

Figure 4: Visualized features from TPSAM.

As shown in Fig. 4, after alignment through SCCM and TPSAM, $A^i_t (i = 2, ..., 4)$ exhibits reduced spatial discrepancies with the RGB image and diminished background noise. This indicates that the common salient regions across the RGB and thermal modalities have been effectively aligned.

Cross-Modal Correlation Module

After the alignment by TPSAM, the salient regions of the warped thermal image A_t^i and the enhanced RGB features $\hat{F}_{rgb}^i$ from SCCM are roughly aligned. However, the correlated cues between the RGB and thermal modalities have not been fully exploited. To address this, the CMCM models the correlation between A_t^i and $\hat{F}_{rgb}^i$ to facilitate feature fusion. As shown in Fig. 5, we employ an efficient scanning mechanism to project features from both modalities into a shared hidden state space. Then, a gating mechanism is used to construct transitions between hidden states, enabling deep cross-modal feature fusion.

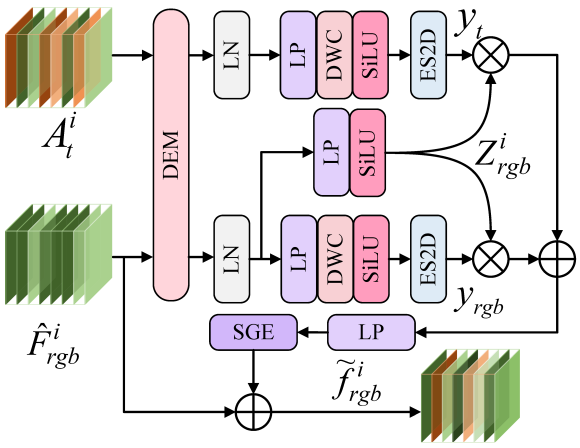

Figure 5: Structure of CMCM (RGB branch).

Specifically, after obtaining the aligned thermal image A_t^i and the enhanced RGB features $\hat{F}_{rgb}^i$, we first enhance their local information using the DEM module. Then, we project both features into a hidden state space through an efficient scanning mechanism, obtaining y_{rgb} and y_t.

$$\begin{aligned} y_{rgb} &= ES2D(SiLU(DWC(LP(LN(DEM(\hat{F}_{rgb}^i)))))) \\ y_t &= ES2D(SiLU(DWC(LP(LN(DEM(A_t^i))))))\end{aligned} \tag{9}$$

The projected features A_t^i and $\hat{F}_{rgb}^i$ are used to generate the gating parameters $Z_{rgb}^i = SiLU(LP(LN(DEM(\hat{F}_{rgb}^i))))$ and $Z_t^i = SiLU(LP(LN(DEM(A_t^i))))$, respectively.

Then, Z_{rgb}^i and Z_t^i are used to modulate y_{rgb} and y_t, enabling cross-modal fusion in the hidden state space and fully leveraging complementary information across branches. The output is then passed through the SGE module to suppress redundant information generated in the hidden state, followed by residual connections to preserve the original information. This process is implemented as follows:

$$\begin{aligned} \tilde{f}_{rgb}^i &= SGE(LP(y_{rgb} \otimes Z_{rgb}^i \oplus y_t \otimes Z_{rgb}^i)) \oplus \hat{F}_{rgb}^i \\ \tilde{f}_t^i &= SGE(LP(y_t \otimes Z_t^i \oplus y_{rgb} \otimes Z_t^i)) \oplus A_t^i \end{aligned} \tag{10}$$

As shown in Fig. 6, the spatial discrepancy between the RGB feature $\tilde{f}_{rgb}^i$ and the thermal feature $\tilde{f}_t^i$ has been significantly reduced after being processed by the CMCM. By fully exploring and integrating both correlation and saliency information, CMCM enables a deeper fusion of bimodal features, resulting in a joint representation that combines the rich texture details of the RGB modality with the strong target perception capabilities of the thermal modality.

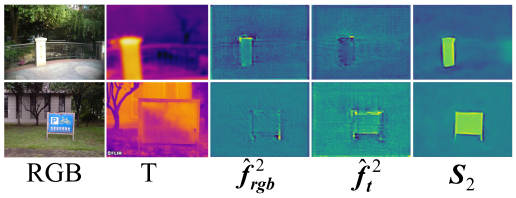

Figure 6: Visualized features from CMCM.

Finally, the strongly correlated features $\tilde{f}_{rgb}^i$ and $\tilde{f}_t^i$ obtained from correlation modeling are fused and decoded. Specifically, $\tilde{f}_{rgb}^i$ and $\tilde{f}_t^i$ are concatenated along the channel dimension and then passed through a 3×3 convolution layer to aggregate their saliency information, as formulated below:

$$S_i = Conv(\left[\tilde{f}_{rgb}^i, \tilde{f}_t^i\right]) \tag{11}$$

where S_i $(i = 2, 3, 4)$ denotes the fused features. The decoder integrates these features in a top-down manner to produce the final prediction, which is supervised by the saliency ground truth and optimized using a combination of binary cross-entropy loss, smoothness loss, and Dice loss.

Experiments

Implementation Details

Our model is implemented on a single RTX 4090 GPU. The model is optimized using the AdamW optimizer with a learning rate of 1e-5, weight decay of 1e-4, a batch size of 4, and trained for 200 epochs. During both training and inference stages, input images are resized to 384×384 pixels.

Datasets and Metrics

For a comprehensive evaluation of the proposed model, we conduct experiments on both unaligned and aligned datasets. Following the methodology in (Wang et al. 2025a), we train our model and comparative methods on the UVT20K training set, and evaluate them on the following test sets: UVT20K (unaligned), UVT2000 (unaligned), un-VT821 (weakly aligned), un-VT1000 (weakly aligned), and un-VT5000 (weakly aligned). Additionally, we train our alignment-based model on the VT5000 training set and test it on the aligned datasets: VT821, VT1000, and VT5000. We employ three widely-used evaluation metrics for comprehensive assessment: E-measure (E_m), S-measure (S_m), and F-measure (F_m).

Comparison with SOTA methods

We conduct comprehensive comparisons between our method and 16 state-of-the-art approaches, including 10 heavyweight RGB-T SOD methods (Table 1) and 6 lightweight ones (Table 2). The compared methods include:

Methods	Backbone	Metrics	UVT20K	UVT2000	un-VT5000	un-VT1000	un-VT821	VT5000	VT1000	VT821	Params(M)	FLOPs(G)
DCNet$_{22}$	VGG16	F_m	0.779	0.621	0.790	0.889	0.799	0.819	0.948	0.823	24.1	246.59
		S_m	0.821	0.770	0.854	0.915	0.860	0.871	0.922	0.876		
		E_m	0.861	0.799	0.908	0.943	0.908	0.920	0.902	0.912		
LAFB$_{24}$	Res2Net50	F_m	0.774	0.620	0.824	0.890	0.792	0.854	0.908	0.842	118.76	139.73
		S_m	0.850	0.789	0.879	0.925	0.860	0.894	0.936	0.892		
		E_m	0.871	0.801	0.912	0.934	0.880	0.928	0.949	0.916		
MSEDNet$_{24}$	ResNet152	F_m	0.508	0.384	0.766	0.841	0.796	0.873	0.919	*0.877*	93.55	111.62
		S_m	0.685	0.651	0.811	0.867	0.839	0.910	0.941	0.917		
		E_m	0.699	0.633	0.901	0.923	0.912	0.943	0.954	0.940		
ConTriNet$_{25}$	Res2Net50	F_m	0.812	*0.694*	0.836	0.894	0.813	0.848	0.899	0.833	34.78	55.42
		S_m	0.852	*0.823*	0.880	0.924	0.872	0.889	0.926	0.883		
		E_m	0.884	0.823	0.922	0.940	0.907	0.889	0.946	0.911		
WaveNet$_{23}$	Wave-MLP	F_m	0.674	0.460	0.664	0.758	0.670	0.864	*0.952*	0.863	80.7	64.02
		S_m	0.792	0.702	0.825	0.875	0.826	0.911	**0.945**	0.912		
		E_m	0.809	0.662	0.831	0.863	0.843	0.940	0.921	0.929		
SPNet$_{23}$	PVT-v2-B3	F_m	0.786	0.639	0.848	0.902	0.833	0.880	0.954	0.873	109.95	56.94
		S_m	0.863	0.808	*0.900*	0.931	0.894	0.914	0.941	*0.913*		
		E_m	0.883	0.803	0.929	0.938	0.910	0.948	0.925	0.936		
SwinNet$_{22}$	SwinB	F_m	0.737	0.579	0.823	0.890	0.799	0.846	0.947	0.818	198.78	124.72
		S_m	0.857	0.800	0.899	*0.936*	0.888	0.912	0.938	0.904		
		E_m	0.844	0.751	0.923	0.938	0.905	0.942	0.894	0.926		
TCINet$_{24}$	SwinB	F_m	*0.832*	0.699	*0.872*	0.907	0.848	0.876	0.909	0.852	88.2	91.87
		S_m	0.842	0.818	0.908	0.934	0.898	0.909	0.935	0.901		
		E_m	0.887	0.825	*0.946*	*0.947*	0.922	0.949	0.949	0.922		
SACNet$_{24}$	SwinB	F_m	0.689	0.594	0.876	0.923	0.857	*0.888*	**0.958**	0.859	300.26	143.78
		S_m	0.841	0.801	**0.910**	0.939	*0.905*	*0.917*	*0.942*	0.906		
		E_m	0.816	0.796	0.949	0.951	*0.929*	0.957	0.927	0.932		
PCNet$_{25}$	SwinB	F_m	0.827	0.691	0.863	*0.910*	0.869	0.899	0.926	0.879	291.91	148.88
		S_m	*0.867*	0.821	0.889	0.922	0.893	0.920	0.939	0.913		
		E_m	*0.894*	*0.830*	0.933	*0.947*	**0.936**	*0.956*	0.946	*0.939*		
TPS-SCL$_{Ours}$	PVT-v2-B4	F_m	0.835	0.699	0.865	0.908	*0.861*	0.884	0.916	0.870	146.85	72.06
		S_m	0.881	0.828	0.904	0.939	**0.911**	0.914	0.940	0.912		
		E_m	0.897	0.831	0.933	0.946	0.927	0.946	*0.950*	0.929		
TPS-SCL$_{Ours}$	SwinB	F_m	**0.848**	**0.702**	**0.893**	**0.924**	**0.874**	**0.902**	0.921	**0.883**	258.26	139.17
		S_m	**0.890**	**0.831**	**0.910**	**0.941**	0.907	**0.922**	0.944	**0.918**		
		E_m	**0.902**	**0.835**	**0.950**	**0.954**	0.934	**0.958**	**0.957**	**0.942**		

Table 1: Comparison with SOTA methods on different datasets. Bold, underlined, italic fonts denote the top 3 methods.

PCNet (Wang et al. 2025a), SACNet (Tu, Qian, and Zhou 2025), TCINet (Lv et al. 2024), SwinNet (Liu et al. 2022), SPNet (Zhang, Wang, and Han 2023), WaveNet (Zhou et al. 2023a), ConTriNet (Tang et al. 2025), MSEDNet (Peng et al. 2024), LAFB (Wang et al. 2024c), DCNet (Tu et al. 2022), LGPNet (Jin et al. 2025), HENet (Gao et al. 2025), LSNet (Zhou et al. 2023b), ORSNet (Huo et al. 2022), MobileSal (Wu et al. 2022), and MoADNet (Jin, Yi, and Xu 2022). To ensure fair comparison, we either use the results reported in the literature or run their publicly available codes with default parameters.

When compared to the lightweight RGB-T SOD methods, TPS-SCL outperforms all competitors, except for a slight performance gap on the VT1000 and un-VT821. As shown in Table 2, compared to the second-best lightweight method, HENet, our TPS-SCL achieves substantial gains for alignment-free datasets. On UVT20K, the gains are +7.0% (F_m), +1.5% (S_m), and +2.9% (E_m), while on UVT2000, the gains are +5.0% (F_m), +1.7% (S_m), and +4.2% (E_m). The Params and FLOPs of our TPS-SCL are 12.82M and 12.34G, respectively.

When it comes to comparison with heavyweight methods, we replace the backbone MobileVit-S with SwinB and PVT-V2-B4. As shown in Table 1, our method (SwinB version) achieves very competitive performance across all eight datasets. Compared to PCNet, our TPS-SCL demonstrates significant improvements. On the largest unaligned dataset (UVT20K), the metrics obtain +2.1% (F_m), +2.3% (S_m), and +0.8% (E_m) gains, while on UVT2000, the gains are +1.1% (F_m), +1.0% (S_m), and +0.5% (E_m). In addition, the PVT-V2 version also demonstrates very competitive performance with much lower Params and FLOPs. For more ablation study on different backbones, please refer to the supplement material on the project website.

The above results demonstrate that our proposed core modules can well adapt to different backbone networks.

Methods	Backbone	Metrics	UVT20K	UVT2000	un-VT5000	un-VT1000	un-VT821	VT5000	VT1000	VT821	Params(M)	FLOPs(G)
MoADNet$_{22}$	MobileNet-V3	F_m	0.237	0.170	0.653	0.772	0.674	0.628	0.745	0.628	5.03	2.96
		S_m	0.483	0.490	0.766	0.828	0.768	0.747	0.810	0.741		
		E_m	0.624	0.618	0.810	0.859	0.809	0.787	0.839	0.781		
MobileSal$_{22}$	MobileNet-V2	F_m	*0.744*	*0.571*	0.705	0.796	0.636	0.713	0.784	0.654	6.55	2.33
		S_m	*0.841*	0.770	0.746	0.796	0.713	0.754	0.803	0.654		
		E_m	*0.849*	0.750	0.808	0.862	0.745	0.812	0.851	0.761		
OSRNet$_{22}$	VGG16	F_m	-	0.454	0.571	0.701	0.575	0.807	0.891	0.801	15.6	-
		S_m	-	0.696	0.724	0.800	0.733	0.875	0.926	0.875		
		E_m	-	0.732	0.770	0.825	0.790	0.908	0.935	0.896		
LSNet$_{23}$	MobileNet-V2	F_m	0.707	0.558	*0.757*	*0.853*	*0.746*	0.827	0.887	0.827	4.57	1.23
		S_m	0.833	0.778	*0.856*	*0.910*	*0.852*	0.876	0.924	0.877		
		E_m	0.831	*0.745*	*0.890*	*0.919*	*0.888*	0.916	0.936	0.911		
HENet$_{25}$	MobileNet-S	F_m	0.745	0.573	0.847	0.893	0.835	0.867	*0.913*	0.867	10.43	10.75
		S_m	0.851	*0.777*	0.894	0.930	0.896	0.900	0.943	0.909		
		E_m	0.858	0.750	0.929	0.939	0.923	0.933	0.950	0.935		
LGPNet$_{25}$	MobileNet-XS	F_m	-	-	-	-	-	*0.850*	0.908	*0.842*	7.35	6.40
		S_m	-	-	-	-	-	*0.890*	*0.934*	*0.890*		
		E_m	-	-	-	-	-	0.908	**0.961**	*0.919*		
TPS-SCL$_{Ours}$	MobileNet-S	F_m	**0.815**	**0.623**	**0.859**	**0.908**	**0.846**	**0.870**	0.915	**0.876**	12.82	12.34
		S_m	**0.866**	**0.794**	**0.896**	**0.934**	0.890	**0.906**	0.944	**0.915**		
		E_m	**0.887**	**0.792**	**0.934**	**0.948**	**0.924**	**0.937**	*0.954*	**0.939**		

Table 2: Comparison with SOTA lightweight methods. Bold, underlined, italic fonts denote the top 3 methods.

Models	UVT20K	UVT2000
TPS-SCL	0.815/0.866/0.887	0.632/0.794/0.792
w/o SCCM	0.022/0.431/0.516	0.024/0.465/0.625
w/o TPSAM	0.625/0.792/0.763	0.498/0.735/0.707
w/o CMCM	0.763/0.804/0.831	0.560/0.710/0.684

Table 3: Ablation study on different components.

Ablation Study

Effectiveness of SCCM We evaluate its impact by removing the SCCM module, which means directly aligning dual-modal features and modeling correlations without high-level semantic constraints. As shown in Table 3, compared to our complete model (TPS-SCL), the average performance drops across three metrics (F_m, S_m, E_m) are 79.3%, 43.5%, and 37.1% on the largest unaligned dataset UVT20K, and 60.8%, 32.9%, and 16.7% on UVT2000, respectively. These results demonstrate that high-level semantic constraints effectively focus RGB and thermal features on salient regions, mitigating misalignment interference. Without SCCM, direct dual-modal alignment introduces significant background noise interference, substantially degrading detection performance.

Effectiveness of TPSAM We remove TPSAM, resulting in unaligned modalities. As shown in Table 3, without the TPSAM module, the average performance drops across three metrics (F_m, S_m, E_m) are 19%, 7.4%, and 12.4% on UVT20K, and 13.4%, 5.9%, and 8.5% on UVT2000, respectively. These results confirm the positive impact of TPSAM in effectively reducing spatial discrepancies between RGB and thermal features.

Effectiveness of CMCM We replace it with direct feature addition, which performs naive multimodal feature fusion. Compared to the complete TPS-SCL model, the version without CMCM ("w/o CMCM") shows degraded performance across all metrics on both unaligned datasets. This conclusively demonstrates that our cross-modal correlation modeling approach is effective for feature fusion.

Conclusion

We propose a TPS-SCL for alignment-free RGB-T SOD. It incorporates both the ES2D and the LSSM to model long-range dependencies. To enhance salient cues in shallow features, we design the SCCM module, which utilizes high-level semantic information to constrain and guide shallow features, thereby providing weakly correlated information with reduced background noise for subsequent multi-modal feature alignment in the TPSAM module. The TPSAM further incorporates a LSSM to strengthen boundary and texture representation capabilities, compensating for potential loss of local neighboring information during efficient scanning. Subsequently, it employs TPS transformation to warp thermal features into co-salient regions of RGB features, effectively mitigating inter-modal spatial discrepancies. We develop the CMCM module to fully exploit inter-modal correlations and complementarity to improve saliency prediction accuracy. Comprehensive experimental results demonstrate that TPS-SCL exhibits superior performance.

Acknowledgments

This work was supported in part by the NSFC under Grant 62472067, the Joint Funds of Liaoning Science and Technology Program under Grant 2023JH2/101800032, and the Taishan Scholar Program of Shandong Province under Grant tstp20221128.

References

Cong, R.; Zhang, K.; Zhang, C.; Zheng, F.; Zhao, Y.; Huang, Q.; and Kwong, S. 2023. Does Thermal Really Always Matter for RGB-T Salient Object Detection? *IEEE Transactions on Multimedia*, 25: 6971–6982.

Duchon, J. 1977. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In Schempp, W.; and Zeller, K., eds., *Constructive Theory of Functions of Several Variables*, 85–100. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-37496-1.

Gao, H.; Wang, F.; Wang, M.; Sun, F.; and Li, H. 2025. Highly Efficient RGB-D Salient Object Detection With Adaptive Fusion and Attention Regulation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(4): 3104–3118.

Huang, T.; Pei, X.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2025. LocalMamba: Visual State Space Model with Windowed Selective Scan. In Del Bue, A.; Canton, C.; Pont-Tuset, J.; and Tommasi, T., eds., *Computer Vision – ECCV 2024 Workshops*, 12–22. Cham: Springer Nature Switzerland. ISBN 978-3-031-91979-4.

Huo, F.; Zhu, X.; Zhang, L.; Liu, Q.; and Shu, Y. 2022. Efficient Context-Guided Stacked Refinement Network for RGB-T Salient Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5): 3111–3124.

Jin, D.; Shao, F.; Xie, Z.; Mu, B.; and Chen, H. 2025. Rethinking Lightweight RGB–Thermal Salient Object Detection With Local and Global Perception Network. *IEEE Internet of Things Journal*, 12(11): 18056–18069.

Jin, X.; Yi, K.; and Xu, J. 2022. MoADNet: Mobile Asymmetric Dual-Stream Networks for Real-Time and Lightweight RGB-D Salient Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7632–7645.

Li, X.; Hu, X.; and Yang, J. 2019. Spatial Group-wise Enhance: Improving Semantic Feature Learning in Convolutional Networks. arXiv:1905.09646.

Liu, Z.; Tan, Y.; He, Q.; and Xiao, Y. 2022. 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.

Lv, C.; Zhou, X.; Wan, B.; Wang, S.; Sun, Y.; Zhang, J.; and Yan, C. 2024. Transformer-Based Cross-Modal Integration Network for RGB-T Salient Object Detection. *IEEE Transactions on Consumer Electronics*, 70(2): 4741–4755.

Mehta, S.; and Rastegari, M. 2022. MobileViT: Lightweight, General-purpose, and Mobile-friendly Vision Transformer. In *International Conference on Learning Representations*.

Peng, D.; Zhou, W.; Pan, J.; and Wang, D. 2024. MSEDNet: Multi-scale fusion and edge-supervised network for RGB-T salient object detection. *Neural Networks*, 171: 410–422.

Song, K.; Wen, H.; Ji, Y.; Xue, X.; Huang, L.; Yan, Y.; and Meng, Q. 2024. SIA: RGB-T salient object detection network with salient-illumination awareness. *Optics and Lasers in Engineering*, 172: 107842.

Tang, H.; Li, Z.; Zhang, D.; He, S.; and Tang, J. 2025. Divide-and-Conquer: Confluent Triple-Flow Network for RGB-T Salient Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3): 1958–1974.

Tu, Z.; Li, Z.; Li, C.; and Tang, J. 2022. Weakly Alignment-Free RGBT Salient Object Detection With Deep Correlation Network. *IEEE Transactions on Image Processing*, 31: 3752–3764.

Tu, Z.; Qian, X.; and Zhou, W. 2025. SACNet: Saliency-Aided Aggregation Consensus Network for RGB-D Co-Salient Object Detection. *IEEE Signal Processing Letters*, 32: 2000–2004.

Wang, J.; Li, G.; Shi, J.; and Xi, J. 2024a. Weighted Guided Optional Fusion Network for RGB-T Salient Object Detection. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(5).

Wang, K.; Chen, K.; Li, C.; Tu, Z.; and Luo, B. 2025a. Alignment-Free RGB-T Salient Object Detection: A Large-Scale Dataset and Progressive Correlation Network . In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7780–7788. Washington, DC, USA: AAAI Press.

Wang, K.; Chen, K.; Li, C.; Tu, Z.; and Luo, B. 2025b. EfficientVMamba: Atrous Selective Scan for Light Weight Visual Mamba. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6443–6451. Washington, DC, USA: AAAI Press.

Wang, K.; Lin, D.; Li, C.; Tu, Z.; and Luo, B. 2024b. Alignment-Free RGBT Salient Object Detection: Semantics-Guided Asymmetric Correlation Network and a Unified Benchmark. *IEEE Transactions on Multimedia*, 26: 10692–10707.

Wang, K.; Tu, Z.; Li, C.; Zhang, C.; and Luo, B. 2024c. Learning Adaptive Fusion Bank for Multi-Modal Salient Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8): 7344–7358.

Wu, Y.-H.; Liu, Y.; Xu, J.; Bian, J.-W.; Gu, Y.-C.; and Cheng, M.-M. 2022. MobileSal: Extremely Efficient RGB-D Salient Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 10261–10269.

Zhang, Z.; Wang, J.; and Han, Y. 2023. Saliency Prototype for RGB-D and RGB-T Salient Object Detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 3696–3705. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Zhou, W.; Sun, F.; Jiang, Q.; Cong, R.; and Hwang, J.-N. 2023a. WaveNet: Wavelet Network With Knowledge Distillation for RGB-T Salient Object Detection. *IEEE Transactions on Image Processing*, 32: 3027–3039.

Zhou, W.; Zhu, Y.; Lei, J.; Yang, R.; and Yu, L. 2023b. LSNet: Lightweight Spatial Boosting Network for Detecting Salient Objects in RGB-Thermal Images. *IEEE Transactions on Image Processing*, 32: 1329–1340.