Alignment-Free RGBT Salient Object Detection: Semantics-guided Asymmetric Correlation Network and A Unified Benchmark

Kunpeng Wang, Danying Lin, Chenglong Li, Zhengzheng Tu, Bin Luo, *Senior Member, IEEE*

***Abstract*—RGB and Thermal (RGBT) Salient Object Detection (SOD) aims to achieve high-quality saliency prediction by exploiting the complementary information of visible and thermal image pairs, which are initially captured in an unaligned manner. However, existing methods are tailored for manually aligned image pairs, which are labor-intensive, and directly applying these methods to original unaligned image pairs could significantly degrade their performance. In this paper, we make the first attempt to address RGBT SOD for initially captured RGB and thermal image pairs without manual alignment. Specifically, we propose a Semantics-guided Asymmetric Correlation Network (SACNet) that consists of two novel components: 1) an asymmetric correlation module utilizing semantics-guided attention to model cross-modal correlations specific to unaligned salient regions; 2) an associated feature sampling module to sample relevant thermal features according to the corresponding RGB features for multi-modal feature integration. In addition, we construct a unified benchmark dataset called UVT2000, containing 2000 RGB and thermal image pairs directly captured from various real-world scenes without any alignment, to facilitate research on alignment-free RGBT SOD. Extensive experiments on both aligned and unaligned datasets demonstrate the effectiveness and superior performance of our method. The dataset and code are available at https://github.com/Angknpng/SACNet.**

***Index Terms*—RGBT salient object detection, alignment-free, asymmetric correlation module, associated feature sampling module.**

I. Introduction

Salient Object Detection (SOD) aims to identify and segment the most attractive regions in visible images. It has been applied in a variety of tasks, such as image retrieval [1], person re-identification [2], object tracking [3], and video analysis [4], [5]. Despite the great success achieved by existing methods [6]–[8], single-modal SOD remains challenging when dealing with deteriorated visible images caused by illumination variation [9], [10], complex background, etc. Since thermal images are captured based on the infrared radiation energy of the scene, they are not interfered by illumination and can capture the overall shape of objects. RGB and Thermal (RGBT) SOD improves the performance of single-modal SOD by introducing the corresponding thermal modality and exploiting their complementary benefits. In practice, the RGBT image pairs captured directly by the device are unaligned. However, existing RGBT SOD datasets are manually aligned, which consumes a lot of labor and fails to reflect the problems caused by unalignment in practical applications. Based on this, existing methods are almost tailored for aligned RGBT image pairs and are difficult to adapt to unaligned cases, which directly leads to performance degradation on unaligned image pairs.

Specifically, existing RGBT SOD methods [11]–[16] integrate the complementary information of the two modalities through different fusion paradigms. Although promising re-

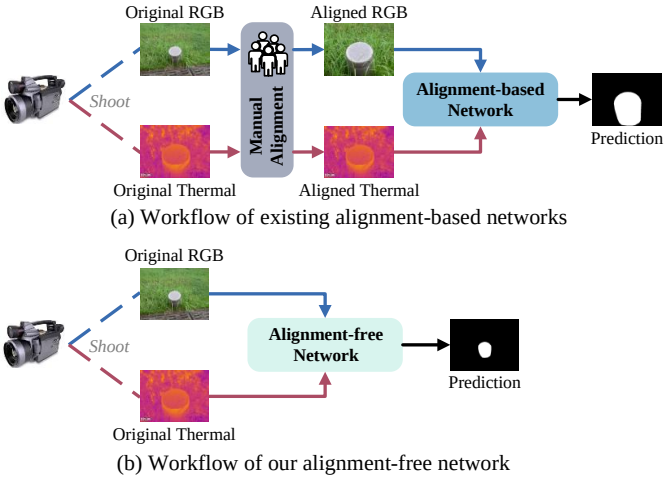


Fig. 1. Workflow comparisons between existing networks and our network. (a) Existing networks require a labor-intensive manual alignment process to align visible and thermal image pairs, further exploiting modality complementarity for saliency prediction. (b) Our network directly mines multi-modal correlations and complementarity of initially captured unaligned image pairs for saliency prediction.

This work is supported in part by University Synergy Innovation Program of Anhui Province (No.GXXT-2022-014), in part by National Natural Science Foundation of China under Grant 62376005, in part by National Natural Science Foundation of China (No. 62376004), in part by Natural Science Foundation of Anhui Province (No. 2208085J18), in part by the Natural Science Foundation of Anhui Higher Education Institution of China under Grant KJ2020A0033, in part by Anhui Provincial Natural Science foundation under Grant 2108085MF211, in part by Anhui Energy Internet Joint Fund Project under Grant 2008085UD07, in part by the National Natural Science Foundation of China under Grant 61876002, in part by Anhui Provincial Key Research and Development Program under Grant 202104d07020008, and in part by the NSFC Key Project of International (Regional) Cooperation and Exchanges under Grant 61860206004. (Corresponding author is Zhengzheng Tu)

Kunpeng Wang, Danying Lin, Zhengzheng Tu and Bin Luo are affiliated with Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: kp.wang@foxmail.com; danying_lin@foxmail.com; zhengzhengahu@163.com and luobin@ahu.edu.cn).

Chenglong Li is affiliated with the Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Security Artificial Intelligence, School of Artificial Intelligence, Anhui University, Hefei 230601, China (e-mail: lcl1314@foxmail.com).

sults are achieved, they are designed upon manually aligned visible and thermal image pairs, which requires heavy labor costs. Fig. 1 (a) illustrates the workflow of existing methods that perform saliency prediction on the basis of manually aligned RGBT image pairs. In the case of alignment, objects in RGB and thermal modalities inherently correspond in space, they are naturally correlated, and simple operations such as summation and concatenation can directly exploit the multi-modal complementarity. However, the initially captured image pairs in practice are unaligned [17], exhibiting deviations in both position and scale, as shown in Fig. 6. In such case, the correlation between the two modalities significantly decreases, making it challenging to directly exploit multi-modal complementarity. Therefore, applying existing methods to initially captured image pairs without alignment can severely degrade their performance.

Recently, DCNet [18] attempts to address this issue for weakly aligned image pairs. It performs random spatial affine transformations on existing aligned datasets [17], [19], [20] to artificially create weakly aligned datasets. Based on this, it models the correlation between weakly aligned image pairs through dynamic convolution and feature-wise affine transformation. Nonetheless, two issues still exist: 1) affine transformation with limited transformation space and dynamic convolution with small receptive field are difficult to deal with large spatial deviations, which exist in the initially captured image pairs; 2) artificially created weakly aligned datasets have small deviations and lack some practical significance. Consequently, DCNet also fails when applied to the directly captured unaligned image pairs. For the first issue, we model the comprehensive correlation between the two modalities through an asymmetric-window based correlation operation, which fully associates the corresponding unaligned multi-modal information. Based on this, we sample and integrate relevant multi-modal features through cascaded deformable convolutions for accurate saliency prediction. For the second issue, we construct an unaligned dataset through real-world camera shooting to facilitate the research of alignment-free RGBT SOD.

To this end, we propose a Semantics-guided Asymmetric Correlation Network (SACNet) to enhance the performance of alignment-free RGBT SOD. Specifically, we propose an Asymmetric Correlation Module (ACM) based on transformer attention [21] to model comprehensive multi-modal correlations. since the two modalities are unaligned in position and scale, the ACM restricts the correlation modeling within a pair of asymmetric windows, which preserves complete corresponding salient regions of the two modalities, thereby reducing the inconsistency interference caused by misalignment. To further reduce the interference of background noise, semantic information is embedded into the ACM to guide the correlation modeling to focus on salient regions. In addition, the AFSM samples relevant thermal features conditioned on RGB features through cascaded deformable convolutions, enabling the integration for corresponding multi-modal information. Guided by the above two modules, our method can model robust correlations between the two modalities to further exploit the multi-modal complementary benefits for accurate saliency prediction. As shown in Fig. 1 (b), our method is able to predict the saliency maps for directly captured unaligned image pairs.

By designing the ACM and AFSM modules, we propose the correlation modeling technique that associates and samples the corresponding information in unaligned RGBT image pairs, which are captured directly in the real world without manual alignment. There are several good impacts and applications for the proposed technique. First, based on this technique, our method achieves alignment-free RGBT SOD, which saves the labor costs caused by manual alignment. Second, the proposed correlation modeling technique can be extended to other multi-modal tasks to enhance the correlation between modalities and improve their performance, such as RGBT tracking [22], RGBT super-resolution [23], and image-text retrieval [24]. Third, the proposed technique provides a basis for the collaborative utilization of multi-modal information. Since manual alignment is not required, the proposed technique can be combined with other techniques and deployed into devices with RGB and thermal sensors (e.g., surveillance cameras, satellites, and UAVs) to achieve intelligent video surveillance, remote sensing image analysis, UAV monitoring, etc.

In addition, we construct a unified benchmark dataset with practical significance, UVT2000, to facilitate research on alignment-free RGBT SOD. UVT2000 contains 2000 unaligned visible-thermal image pairs with ground truth annotations, directly captured by a pair of thermal infrared and CCD cameras without any alignment. Therefore, the misalignment of the image pairs is a natural result of camera shooting and reflects the issues in practical applications. Additionally, the image pairs are collected from a variety of real-world scenarios, which are annotated with 11 challenge attributes to facilitate the study of specific issues.

To the best of our knowledge, this is the first work to launch the alignment-free setting and the corresponding benchmark dataset for RGBT SOD. The main contributions of our work are as follows:

- For the first time, we perform RGBT SOD on initially captured visible-thermal image pairs without any manual alignment, which can significantly reduce labor costs.
- We propose an Asymmetric Correlation Module (ACM) to model multi-modal correlations specific to salient regions, and an Associated Feature Sampling Module (AFSM) to sample and integrate relevant features of the two modalities.
- We construct a novel benchmark dataset, containing 2000 unaligned visible-thermal image pairs directly captured from various real-word scenes, to facilitate research on alignment-free RGBT SOD.
- Our proposed method achieves state-of-the-art performance on both aligned and unaligned datasets, demonstrating its effectiveness and the potential of alignment-free RGBT SOD.

II. Related Work

A. RGBD Salient Object Detection

In the past decades, a large number of Salient Object Detection (SOD) methods [25] have been developed through feature

refinement [26]–[28], attention mechanism [29], boundary enhancement [30], [31], uncertainty perception [32], [33], etc. However, they still struggle to handle some challenging scenes such as similar foreground and background, low illumination, and image clutter. To address these issues, some studies [34], [35] initially introduce depth maps with spatial structure information to enhance the performance of single-modal SOD, called RGBD SOD. In order to exploit the multi-modal complementary information, existing methods [36] mainly fuse visible images and depth maps through early fusion, middle fusion, and late fusion.

Early fusion integrates visible images and depth maps into a joint representation as input to a network. For example, Qu et al. [35] compute the joint prior features of visible images and depth maps as input to a convolutional neural network to extract a unified multi-modal feature representation. Song et al. [37] predict multi-level saliency maps based on multi-scale pre-segmentation results of the input RGBD image. Middle fusion mainly utilizes multi-scale feature fusion strategies to mine multi-modal correlations. BBSNet [38] divides multi-level features into teacher and student features, and utilizes the discriminative semantics of the teacher features to suppress the interference in the student features. Cong et al. [39] effectively utilize multi-modal information by progressively integrating multi-level features in the encoder and decoder based on the attention mechanism. Wen et al. [40] improve the universality and anti-interference of saliency predictions by enhancing extracted features and inferring high-level semantic information. Late fusion learns high-level features or saliency maps of the two modalities for fusion. Han et al. [41] learn the feature representations of RGB and depth modalities separately, and mine their complementary relationships to obtain a joint representation for saliency prediction. Zhang et al. [42] independently enhance intra-modal features of two modalities and then selectively interact them based on scene information.

Although these methods exploit the complementary information of the RGB and depth modalities through different fusion strategies, they are designed on two well-aligned modalities, making it difficult to transfer to unaligned multi-modal inputs for correlation modeling.

B. RGBT Salient Object Detection

Since thermal images are captured based on the infrared radiation of objects without suffering from the interference of complex environments (e.g., illumination and fog), several recent researches [17], [19], [20] introduce the thermal modality based on the RGB modality to form RGBT SOD. Existing methods mainly focus on mining modality complementarity [13]–[15], [43], alleviating the modality gap [44], [45], or addressing modality-specific challenges such as low illumination [12], [46], [47] and thermal crossover [48], etc.

For example, to handle the challenge of low illumination, Zhang et al. [12] integrate multi-level and multi-modal features, Liao et al. [48] facilitate the multi-modal interaction in the encoder for discriminative feature extraction. To exploit modality complementarity for accurate prediction, Tu et al. [13] design a dual-decoder and a multi-type interaction. Wang et al [14] excavate multi-modal complementary information with unique single-modal information for mutual guidance. Liu et al. [15] utilize the strong feature representation capability of Swin Transformer to alleviate the gap between the two modalities. For practical applications, Zhou et al. [49] begin to reduce computational complexity with lightweight operations. Although these methods have achieved success in addressing different issues, they all perform on manually aligned RGBT image pairs, incurring expensive labor annotation costs.

Recently, Tu et al. [18] initially address this issue for artificially weakly aligned image pairs by employing affine transformation and dynamic convolution. Although shown to be effective, the limited transform space of affine transformation and the small receptive field of dynamic convolution limit the correlation modeling for large spacial deviations. Additionally, the artificially created datasets lack practical significance. Different from existing models, we directly address RGBT SOD for original captured visible and thermal image pairs without any manual alignment, and construct a novel unaligned dataset to validate the effectiveness of our method.

C. Attention-based Methods

Since attention mechanism [21], [50] has powerful global context modeling capabilities, they have been applied to various fields, such as visual tracking [51], object detection [52], semantic segmentation [53], and image super-resolution [54], demonstrating competitive performance. For example, Carion et al. [52] regard the object detection task as a set prediction problem and build a transformer encoder-decoder framework for detection that eliminates unnecessary manually designed components in the detection process. Zheng et al. [53] apply the attention mechanism to the segmentation field pioneeringly. They use a vision transformer [50] to extract hierarchical features, which are fed into a decoder to obtain predictions. Chen et al. [51] design a transformer tracker that uses transformer attention to establish associations between template and search region features to highlight useful target information.

Some recent methods [15], [55]–[57] have also shown the effectiveness of attention mechanism in SOD tasks. For example, Liu et al. [55] construct the first model based on a pure transformer architecture for RGB and RGBD SOD tasks. By introducing task-related tokens and patch-task-attention in the decoder, boundary and saliency prediction are jointly implemented. Tang et al. [57] fuse the information of two modalities at the input level and send it into a high-resolution transformer, which can maintain high resolution and preserve large receptive fields, to interact multi-modal features for saliency prediction. Pang et al. [56] improve the computation efficiency of transformer attention by aggregating and converting pixel-level tokens to patch-level tokens before the multiplication operation of attention. Based on it, an efficient top-down transformer-based information propagation strategy is designed to integrate multi-modal features.

Nevertheless, to the best of our knowledge, there is no attention-based method designed to model the correlation of salient objects in unaligned image pairs. In this paper, we propose a semantics-guided asymmetric attention to model unaligned multi-modal correlations.

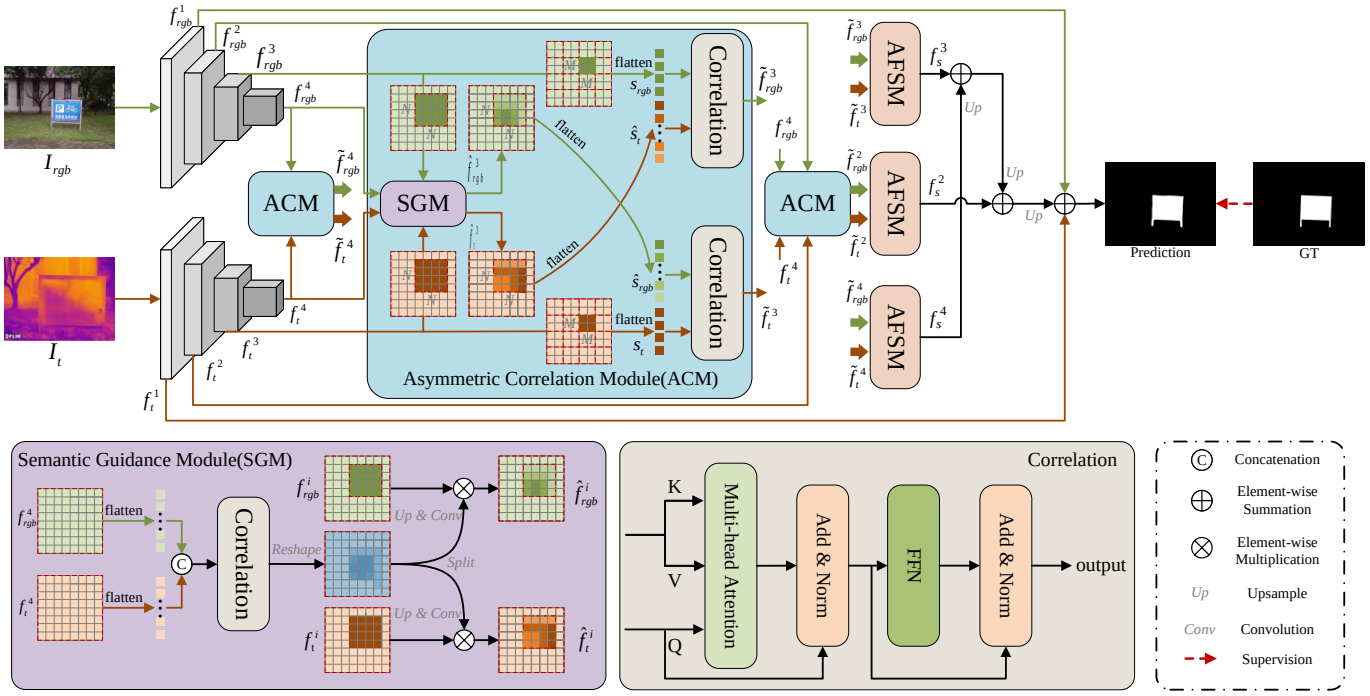


Fig. 2. The overall architecture of our proposed SACNet. The framework mainly comprises an Asymmetric Correlation Module (ACM) and an Associated Feature Sampling Module (AFSM). The ACM restricts the correlation operation within asymmetric window pairs to model comprehensive correlations of the two unaligned modalities. With the Semantic Guidance Module (SGM), ACM focus more on salient regions. In the AFSM, relevant thermal features are sampled according to corresponding RGB features. Subsequently, the multi-modal saliency cues are integrated complementarily for saliency prediction.

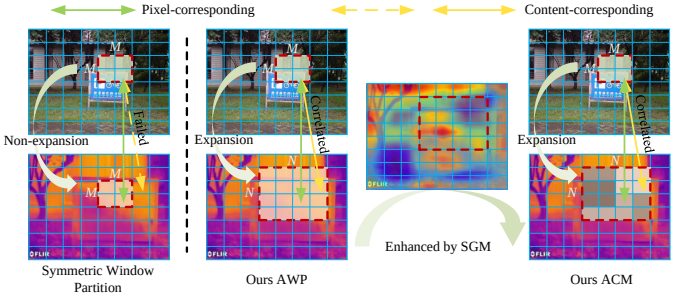


Fig. 3. Comparison of correlation modeling in unaligned RGBT image pairs between the conventional symmetric window partition and the proposed Asymmetric Correlation Module (ACM), which contains the Asymmetric Window Partition (AWP) strategy and Semantic Guidance Module (SGM).

III. Our Method: SACNet

In this paper, we propose a Semantics-guided Asymmetric Correlation Network (SACNet) to model strong multi-modal correlations, thereby leveraging the complementary information of unaligned image pairs for saliency prediction. In the following, we present an overview of SCANet in Section III-A, describe the details of ACM and AFSM in Sections III-B and III-C, respectively, and formulate the loss function in Section III-D.

A. Overview

In this section, we describe the pipeline of the proposed method for alignment-free RGB and Thermal (RGBT) salient object detection (SOD). The overall architecture of our framework is shown in Fig. 2, which consists of two parallel backbones for multi-modal feature extraction, an Asymmetric Correlation Module (ACM), and an Associated Feature Sampling Module (AFSM). The multi-level features extracted by the backbones are collected and denoted as f_m^i $(m \in \{rgb, t\}, i = 1, ..., 4)$. As the salient objects in unaligned image pairs have inconsistent positions and scales, the ACM exploits asymmetric window attention with semantic guidance to model comprehensive correlations between the modalities. Then, the AFSM based on cascaded deformable convolutions is proposed to sample and integrate relevant multi-modal features.

B. Asymmetric Correlation Module (ACM)

Existing RGBT Salient Object Detection methods [13], [14], [16], [45] commonly design different fusion schemes to explore the multi-modal complementarity between aligned image pairs for accurate prediction. Salient regions in aligned image pairs are consistent in spatial location and scale, with strong modality correlation, which facilitates further multi-modal feature integration. Nevertheless, the manual alignment of initially captured unaligned image pairs incurs substantial labor requirements. Furthermore, directly applying these methods to the unaligned image pairs may result in performance degradation, as they struggle to effectively model multi-modal correlations without prior alignment.

In order to establish robust multi-modal correlations between unaligned image pairs, we propose the Asymmetric Correlation Module (ACM), as shown in Fig. 2. Due to the different object positions and scales between the two unaligned modalities, the multi-modal information does not correspond

spatially. Thus, dividing the feature maps of the two unaligned modalities with symmetric windows fails to cover complete correspondence information, resulting in insufficient correlation modeling. Moreover, directly modeling correlations over entire feature maps introduces too much undesired background noise. To this end, the ACM divides the feature map of one modality into small windows, and expands each of these windows into a large one on the feature map of another modality to cover the complete corresponding information. In addition, semantic information is introduced to guide the correlation operation within each asymmetric window pair to focus on object information. In specific, the ACM contains Asymmetric Window Partition (AWP) to alleviate the interference of spatial inconsistency on correlation modeling, and Semantic Guidance Module (SGM) to focus the correlation modeling on salient regions. Fig. 3 vividly demonstrates that ACM can establish sufficient multi-modal correlation for unaligned RGBT image pairs compared to the symmetric window partition.

1) Asymmetric Window Partition (AWP): As shown in Fig. 6, the salient regions in unaligned image pairs have different scales, resulting in different background contents and proportions. In this case, directly applying the original correlation operation across the entire multi-modal regions might introduce excessive undesired background noise, thereby confusing the correlation modeling for salient regions. To mitigate this concern, we first restrict the correlation operation within a window pair. Then, considering the unknown position shifts and scale differences, we adjust the window pair shape to be asymmetric to incorporate complete corresponding information. Although the inconsistent background noise is not completely eliminated in the asymmetric window pair, due to the relatively small window space and the powerful relationship modeling ability of the correlation operation, the AWP is able to establish sufficient multi-modal correlations with less background noise interference.

To be specific, for each modality feature map $f_m^i \in \mathbb{R}^{C\times H\times W}$, we first partition it with a small pixel-level window of size M × M. Due to the spatial inconsistency, we divide the feature map of the other modality with a larger N × N window. Each larger window is obtained by expanding around the small window, aiming to cover the complete corresponding information. In this way, the feature maps with the small window are partitioned into $\lceil \frac{H}{M} \rceil \times \lceil \frac{W}{M} \rceil$ non-overlapping windows. Since the position of the large window is determined by the corresponding small window, the number of large windows in the feature map is also $\lceil \frac{H}{M} \rceil \times \lceil \frac{W}{M} \rceil$ but overlapped. Subsequently, the features within each large window are used to enhance those within the corresponding small one. For example, as shown in Fig. 2, given a RGB feature map size of 8 × 8 with a window size of 2 × 2, the number of windows would be 4 × 4. To preserve the spacial consistency, the window size for the corresponding thermal feature map is 4 × 4, which is centered on the small RGB window. Note that we perform the above operations on both modality features to ensure that each modality is enhanced by the asymmetric correlation, thus achieving a bi-directional cross-modal correlation.

2) Semantic Guidance Module (SGM): To focus cross-modal correlations on salient regions, we propose the SGM, which embeds semantic guidance into the correlation modeling. Since high-level features inherently contain rich semantic information that is capable of object category recognition and localization [58], we utilize the high-level features of both modalities to enhance multi-level feature representations. However, the original high-level features of the two modalities also lack strong correlations and their modality complementarity is not explored, hindering accurate salient region localization. Therefore, the multi-modal high-level features need to be fully aggregated first. As shown at the bottom left of Fig. 2, we collect the top extracted features (f_{rgb}^4 and f_t^4) as high-level features. Considering that the top features have large receptive fields and contain less noise, we directly use the correlation operation on their concatenated features. In this way, the intra-modal and inter-modal correlations can be established simultaneously to obtain a global multi-modal feature representation f_{cat}^G, which is formulated as:

$$f_{cat}^4 = [\text{flatten}(f_{rgb}^4), \text{flatten}(f_t^4)], \tag{1}$$

$$f_{cat}^G = \mathcal{C}(f_{cat}^4, f_{cat}^4), \tag{2}$$

where $[,]$ represents the concatenation operation, and f_{cat}^4 is the high-level concatenated feature. $\mathcal{C}$ is the correlation operation:

$$C(Q,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V+Q, \tag{3}$$

where $Q \in \mathbb{R}^{N_Q\times d_k}$, K, $V \in \mathbb{R}^{N_V\times d_k}$, and '+' represents the residual connection. The details can be seen at the bottom right of Fig. 2, and we refer the reader to the literature [21] for more detailed descriptions. Next, we split f_{cat}^G along the spatial dimension into the global features of the two modalities, and reshape them into their initial shapes, denoted as f_{rgb}^G and f_t^G. To match the spatial and channel dimensions of the features in different layers, we upsample them followed by a 3 × 3 convolutional layer. By multiplying the global feature with the feature maps to be partitioned, the salient regions in each window can be enhanced. This process can be formulated as:

$$\hat{f}_{rgb}^i = Conv(Up_{2^{4-i}}(f_{rgb}^G)) \odot f_{rgb}^i \tag{4}$$

$$\hat{f}_t^i = Conv(Up_{2^{4-i}}(f_t^G)) \odot f_t^i, \tag{5}$$

where $\hat{f}_{rgb}^i$ and $\hat{f}_t^i$ denote the enhanced features, $\odot$ is element-wise multiplication, $Conv$ is a convolution layer, Up_x is the $x\times$ upsample operation with bilinear interpolation, and $i = 2, ...4$ means that we will not perform the correlation operation on the first-level features due to the heavy computational cost of high-resolution data.

It is assumed that sequences from the small window of original RGB and thermal features are denoted as $s_{rgb} = [s_{rgb}^1, s_{rgb}^2, ..., s_{rgb}^{M\times M}]$ and $s_t = [s_t^1, s_t^2, ..., s_t^{M\times M}]$, and sequences from the corresponding large window of enhanced RGB and thermal features are denoted as $\hat{s}_{rgb} = [\hat{s}_{rgb}^1, \hat{s}_{rgb}^2, ..., \hat{s}_{rgb}^{N\times N}]$ and $\hat{s}_t = [\hat{s}_t^1, \hat{s}_t^2, ..., \hat{s}_t^{N\times N}]$. Therefore,

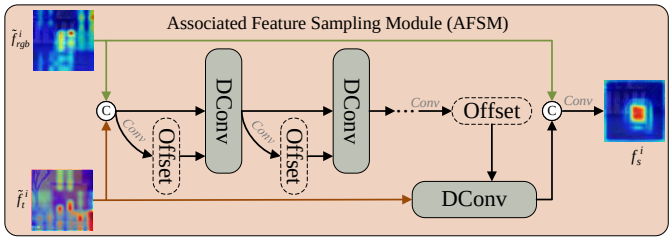


Fig. 4. Details of the proposed associated feature sampling module

the modality correlation modeled by ACM can be formulated as:

$$\begin{aligned} y_{rgb} &= C(s_{rgb}, \hat{s}_t) \\ y_t &= C(s_t, \hat{s}_{rgb}), \end{aligned} \tag{6}$$

where y_{rgb} and y_t denote the output correlated sequences of the two modalities. M and N are the small window size and large window size, respectively. Through correlation operations on all asymmetric window pairs, strong cross-modal correlations for the whole salient region are established, and the feature representations of two modalities are thus improved, denoted as $\tilde{f}_{rgb}^i$ and $\tilde{f}_t^i$.

C. Associated Feature Sampling Module (AFSM)

The features of the two modalities are correlated and enhanced by the ACM, but remain unaligned. Directly aggregating the unaligned features may lead to mismatches and inaccurate recognition for salient regions. To address this issue, we propose an Associated Feature Sampling Module (AFSM), which employs cascaded deformable convolutions (DConvs) [59] to sample relevant multi-modal features for further integration. The architecture of AFSM is shown in Fig. 4. Here, we sample thermal features conditioned on corresponding RGB features. The AFSM first takes both RGB feature $\tilde{f}_{rgb}^i$ and thermal feature $\tilde{f}_t^i$ as inputs to predict sampling offsets α^i for the thermal feature $\tilde{f}_t^i$:

$$\alpha^i = f_\alpha^i(\tilde{f}_t^i, \tilde{f}_{rgb}^i) = \{\Delta p_n | n = 1, ..., |R|\}, \tag{7}$$

where $R = \{(-1,-1),(-1,0),...,(0,1),(1,1)\}$ represents a standard grid of a 3 × 3 kernel, f_α^i denotes the offset generation function. Next, we take α^i and $\tilde{f}_t^i$ as inputs for deformable convolution to compute a newly sampled thermal feature $\tilde{f}_{t\to rgb}^i$. Each position p_0 on the sampled thermal feature map can be formulated as:

$$\tilde{f}_{t\to rgb}^i(p_0) = \sum_{p_n \in R} \omega(p_n) \tilde{f}_t^i(p_0 + p_n + \Delta p_n), \tag{8}$$

where Δp_n is a learnable offset, which may be fractional and thus computed through bilinear interpolation, more details can be found in [59]. In practice, our AFSM cascades four deformable convolutional layers to sample more relevant and accurate saliency features in the thermal modality. Then, the sampled thermal features and corresponding RGB features are concatenated along the channel dimension, and fed into a convolutional layer with a 3 × 3 kernel to aggregate their saliency information, as follows:

$$f_s^i = Conv([\tilde{f}_{t\to rgb}^i, \tilde{f}_{rgb}^i]), \tag{9}$$

where f_s^i $(i = 2, ..., 4)$ denotes the integrated multi-modal features, which will be aggregated for final prediction.

D. Saliency Prediction and Loss Function

Similar to U-Net framework [63], the decoding process for saliency map prediction is formed through top-down feature fusion, which can be formulated as:

$$\hat{f}_s^i = \begin{cases} Conv(Up_2(f_s^i + \hat{f}_s^{i+1}), i = 2, 3 \\ Conv(Up_2(f_s^i), i = 4 \end{cases} \tag{10}$$

$$S = Conv(Up_4(Conv(\hat{f}_s^2 + f_{rgb}^1 + f_t^1))), \tag{11}$$

where S denotes the final predicted saliency map. Note that due to the large computational cost of high-resolution feature maps, the first-level features (f_{rgb}^1 and f_t^1) are only introduced here to supplement detailed information.

Following [13], [64], we use a combination of binary cross-entropy loss, smoothness loss [65], and dice loss [66] to optimize the proposed method, which can be described as:

$$\mathcal{L} = \mathcal{L}_{bce}(S, G) + \mathcal{L}_{smooth}(S, G) + \mathcal{L}_{dice}(S, G), \tag{12}$$

where $\mathcal{L}$ represents the loss function, and G denotes the ground truth.

IV. UVT2000 Benchmark

Existing RGBT SOD datasets [17], [19], [20] are manually aligned, which is labor-intensive and limits the research on alignment-free RGBT SOD. Although DCNet [18] builds weakly aligned datasets (i.e., unaligned-VT5000 [18], unaligned-VT1000 [18], unaligned-VT821 [18]) by performing random affine transformation for the image pairs in existing datasets (i.e., VT5000, VT1000, and VT821), the misalignment of these image pairs is artificial and weak with a lack of practical significance. To facilitate the research on alignment-free RGBT SOD, we construct a benchmark dataset with 2000 unaligned visible-thermal image pairs, denoted as UVT2000.

A. Data Acquisition

UVT2000 dataset is captured by a FLIR SC620 equipped with a pair of CCD and thermal infrared cameras, the same as used in [17]. Due to parallax and different size viewing angles between the two cameras, the same object in their captured image pairs inherently suffers from positional shifts and scale differences, as shown in Fig. 6. Therefore, the degree of misalignment in UVT2000 derives from actual camera shots, and reflects the practical situations in real-world applications. UVT2000 also avoids a number of labor-intensive operations such as manual cropping and rescaling.

Given that visible images are more in line with human visual preferences, we capture these image pairs mainly based on

TABLE I
COMPARISON OF UVT2000 WITH THE PREVALENT MULTI-MODAL SOD DATASETS ON MODALITY, SCENE NUMBER, CATEGORY NUMBER, COLLECTION METHOD, MANUAL ALIGNMENT, CHALLENGE ANNOTATION, AND RESOLUTION.

	Dataset	Year	Modality			Scene	Category	Collection	Manual	Challenge	Resolution		
			RGB	Thermal	Depth	Number	Number	Method	Alignment	Annotation	RGB	Thermal	Depth
Aligned	VT821 [19]	2018	821	821	-	16	198	Camera	✓	✓	640 × 480	640 × 480	-
	VT1000 [17]	2019	1000	1000	-	76	329	Camera	✓	✓	640 × 480	640 × 480	-
	VT5000 [20]	2020	5000	5000	-	212	304	Camera	✓	✓	640 × 480	640 × 480	-
	NJUD [60]	2014	1985	-	1985	282	359	Movie/Internet	✓	×	256 × 256	-	256 × 256
	DUT-RGBD [61]	2019	1200	-	1200	191	291	Camera	✓	×	256 × 256	-	256 × 256
	SIP [62]	2020	929	-	929	69	1	Mobile phone	✓	×	256 × 256	-	256 × 256
Weakly Aligned	un-VT821 [18]	2022	821	821	-	16	198	Augmentation of existing datasets	✓	✓	640 × 480	640 × 480	-
	un-VT1000 [18]	2022	1000	1000	-	76	329		✓	✓	640 × 480	640 × 480	-
	un-VT5000 [18]	2022	5000	5000	-	212	304		✓	✓	640 × 480	640 × 480	-
Unaligned	UVT2000	-	2000	2000	-	295	429	Camera	×	✓	2048 × 1536	640 × 480	-

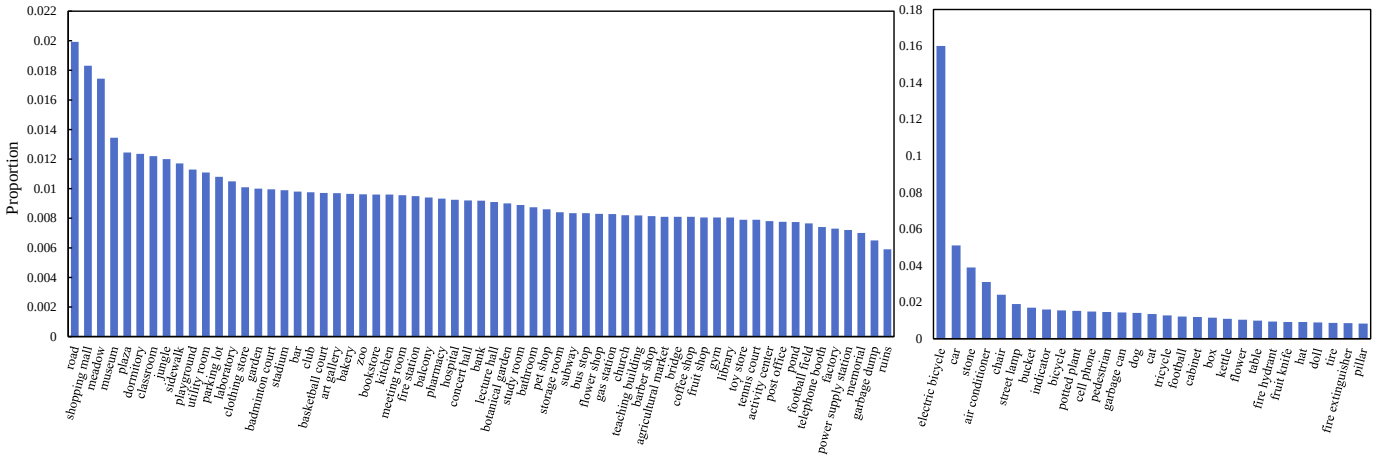


Fig. 5. Top 60% scene and object category distributions in our proposed UVT2000 dataset.

RGB modality, except for visible image degradation cases, such as low illumination and out of focus. To ensure the quality and reliability of the dataset, we initially collect about 2500 pairs of images for selection.

B. Dataset Annotation

Before annotation, we first discard obviously low-quality image pairs, such as image pairs with no objects. Then, five annotators are asked to select the objects in each image pair that are salient at first glance. Based on the agreement of the selected salient objects, we rank these image pairs. Eventually, we annotate the highest-ranking 2000 image pairs. Since the data collection process is mainly based on the RGB modality, we annotate pixel-level ground truth according to visible images supplemented by corresponding thermal images.

The image pairs in our dataset are captured in various scenes. We annotate them with 11 challenges to facilitate different methods to address these challenges. These challenges are: big salient object (BSO), small salient object (SSO), low illumination (LI), bad weather (BW), multiple salient objects (MSO), center bias (CB), cross image boundary (CIB), similar appearance (SA), thermal crossover (TC), image clutter (IC), and out of focus (OF). More detailed descriptions of these challenges can be found in [17]. Fig. 6 shows the examples under different challenges. Note that each image pair may be annotated with multiple challenge attributes, we only describe the main challenge for each example here. Fig. 7 also illustrates the challenge distribution of the UVT2000.

In particular, the image pairs of each scene in Fig. 6 appear to suffer from center bias, but center bias is not a common challenge in the UVT2000 dataset for the following reasons. First, although objects in the thermal modality are generally shifted from the image center due to unalignment, the challenge annotations are mainly based on the RGB modality, which is not always shifted from the image center. Second, as defined in [17], center bias refers to the center of salient objects being far away from the image center. Although the salient objects in some examples appear to be close to the image boundary, their centers are still close to the image center, so they are not grouped into the center bias challenge, such as the examples in the seventh and ninth columns in Fig. 6.

C. Dataset Characteristics

As shown in Table I, compared with existing prevalent multi-modal SOD datasets, UVT2000 mainly has the following characteristics:

1) **No manual alignment is required for constructing UVT2000.** Since UVT2000 is proposed to facilitate the study of alignment-free RGBT SOD, the image pairs are captured directly by the camera without any manual alignment operations.

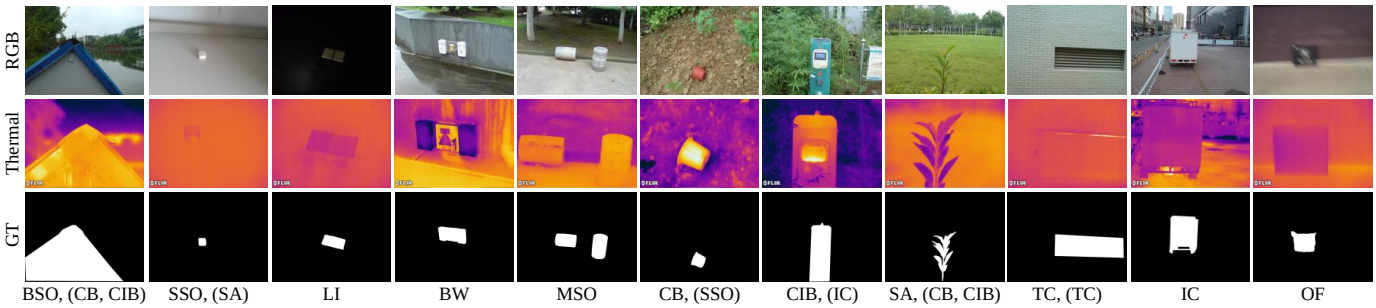


Fig. 6. Examples of visible-thermal image pairs with ground truth and challenge annotations in the proposed benchmark dataset UVT2000. GT: ground truth; BSO: big salient object; SSO: small salient object; LI: low illumination; BW: bad weather; MSO: multiple salient objects; CB: center bias; CIB: cross image boundary; SA: similar appearance; TC: thermal crossover; IC: image clutter; OF: out of focus. A, (B) indicates that the image pair contains the main challenge A and other secondary challenge(s) B.

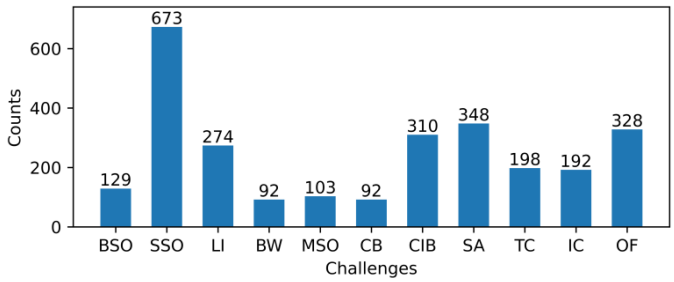


Fig. 7. Challenge distribution of the newly constructed dataset UVT2000.

2) **The UVT2000 is more diverse in scenes and categories.** Based on [67], Table I reports the total number of scenes and categories for UVT2000 and the compared datasets. The results show that UVT2000 has more diverse scenes and categories. Fig. 5 further illustrates the distribution of the top 60% scenes and object categories in the proposed UVT2000. It can be seen that most scenes and categories have approximately smooth distributions.

3) **UVT2000 is more challenging and practical.** As the first unaligned RGBT SOD dataset, compared to existing RGBT SOD datasets, UVT2000 is of smaller scale than VT5000 and unaligned-VT5000 (i.e., a variant of VT5000), while it is more consistent with actual scenes and presents greater challenges. As shown in Table I, compared with the weakly aligned datasets, UVT2000 is collected through real-world camera shooting, which has more practical significance.

4) **High resolution of visible images is maintained in UVT2000.** With no rescaling and cropping required for alignment, the UVT2000 maintains the original resolution of image pairs with high quality.

V. Experiments

A. Datasets

To evaluate the effectiveness of our method, we conduct experiments on three aligned datasets, three weakly aligned datasets, and one unaligned dataset, including VT821 [19], VT1000 [17], VT5000 [20], unaligned-VT821 [18] (i.e., un-VT821), unaligned-VT1000 [18] (i.e., un-VT1000), unaligned-VT5000 [18] (i.e., un-VT5000), and the proposed UVT2000. VT821 contains 821 aligned image pairs, some of which are added with noise to make the dataset more challenging. VT1000 includes 1000 pairs of visible-thermal images that are aligned and collected in relatively simple scenes. VT5000 consists of 5000 aligned image pairs with a variety of object sizes and scenes, which are divided in half into a training set and a testing set. The un-VT821, un-VT1000, and un-VT5000 are weakly aligned datasets obtained by performing random affine transformations on the corresponding aligned datasets above. Following [18], we utilize the training sets of VT5000 to train the model for aligned datasets, and train the model with the training sets of un-VT5000 for weakly aligned and unaligned datasets. The remaining datasets are used for testing.

B. Evaluation Metrics

We adopt four widely used evaluation metrics for evaluation, including E-measure (E_ξ), S-measure (S_α), weighed F-measure (F_β^ω), and mean absolute error (MAE). Specifically, E-measure measures both image-level statistics and pixel-level matching information. S-measure evaluates the structural similarity at the region and object level. Weighed F-measure is a weighted combination of precision and recall. Mean absolute error indicates the absolute error between predictions and ground truth. We also introduce the "precision-recall" curve to demonstrate the overall performance of the model. To assess model complexity, we also report the results in terms of FPS (Frame-Per-Second) and number of parameters.

C. Implementation Details

Our framework is implemented with Pytorch in a workspace with two RTX 3090 GPUs. Input images are resized into 384 × 384 for both training and testing. During the training stage, we apply the AdamW algorithm with a learning rate of 1e-5 and a weight decay of 1e-4 to optimize our network. We set the batch size as 8 and epoch as 200 to train all our models, which takes about 15 hours. The backbone of our network is the SwinB network [73] pre-trained on ImageNet. In our network, the small window size and large window size are set to 4 and 6, respectively.

D. Comparison with State-of-the-arts

We compare our method with 14 state-of-the-art RGBT SOD methods, including ADF [20], MIDD [13], CSRNet [68],

TABLE II
QUANTITATIVE COMPARISON OF E-MEASURE (E_ξ), S-MEASURE (S_α), WEIGHED F-MEASURE (F_β^ω), AND MEAN ABSOLUTE ERROR (MAE) ON ONE UNALIGNED, THREE WEAKLY ALIGNED, AND THREE ALIGNED DATASETS. THE BEST THREE RESULTS ARE MARKED WITH **RED**, **GREEN**, AND **BLUE**.

	Method	ADF$_{20}$ [20]	MIDD$_{21}$ [13]	CSRNet$_{21}$ [68]	CGFNet$_{21}$ [14]	SwinNet$_{22}$ [15]	OSRNet$_{22}$ [69]	TNet$_{22}$ [70]	DCNet$_{22}$ [18]	MCFNet$_{23}$ [71]	HRTransNet$_{23}$ [57]	LSNet$_{23}$ [49]	CAVER$_{23}$ [56]	WaveNet$_{23}$ [16]	SPNet$_{23}$ [72]	SACNet Ours	SACNet Ours
	Backbone	VGG16	VGG16	ESPNetv2	VGG16	SwinB	VGG16	ResNet50	VGG16	ResNet50	HRFormer	MobileNet-v2	ResNet50	Wave-MLP	PVT-v2	ResNet50	SwinB
	FPS ↑	27	33	75	18	34	**142**	**81**	43	72	37	**314**	67	17	29	19	27
	Parameters(M) ↓	66.8	52.4	**1.0**	66.4	199.2	**15.6**	87.0	24.1	70.8	26.3	**4.6**	55.8	80.7	110.0	530.9	327.7
UVT2000	E_ξ ↑	0.640	0.673	0.605	0.705	**0.743**	0.732	0.717	0.753	0.727	0.603	0.679	0.727	0.634	0.733	**0.777**	**0.812**
	S_α ↑	0.672	0.721	0.629	0.740	**0.777**	0.696	0.754	0.731	0.739	0.651	0.728	0.749	0.699	**0.765**	0.759	**0.807**
	F_β^ω ↑	0.349	0.445	0.349	0.477	**0.551**	0.454	0.527	0.513	0.464	0.375	0.466	0.527	0.399	**0.558**	0.544	**0.640**
	MAE ↓	0.078	0.075	0.116	0.067	**0.051**	0.051	0.058	0.049	0.066	0.116	0.065	0.061	0.078	0.059	**0.038**	**0.036**
VT5000	E_ξ ↑	0.891	0.897	0.905	0.922	0.942	0.908	0.927	0.920	0.924	**0.945**	0.915	0.924	0.940	**0.948**	0.933	**0.957**
	S_α ↑	0.864	0.868	0.868	0.883	**0.912**	0.875	0.895	0.871	0.887	**0.912**	0.877	0.892	0.911	**0.914**	0.892	**0.917**
	F_β^ω ↑	0.722	0.763	0.796	0.831	0.846	0.807	0.840	0.819	0.836	**0.870**	0.806	0.835	0.864	**0.880**	0.838	**0.888**
	MAE ↓	0.048	0.043	0.042	0.035	0.026	0.040	0.033	0.035	0.033	**0.025**	0.037	0.032	0.026	**0.024**	0.030	**0.021**
un-VT5000	E_ξ ↑	0.824	0.885	0.804	0.899	0.923	0.770	0.910	0.908	0.905	0.847	0.890	0.917	0.831	**0.929**	**0.925**	**0.949**
	S_α ↑	0.813	0.854	0.746	0.865	**0.899**	0.724	0.879	0.854	0.864	0.811	0.856	0.884	0.825	**0.900**	0.879	**0.911**
	F_β^ω ↑	0.625	0.740	0.602	0.746	**0.823**	0.571	0.806	0.790	0.757	0.692	0.757	0.822	0.664	**0.848**	0.820	**0.876**
	MAE ↓	0.072	0.049	0.089	0.046	**0.031**	0.106	0.038	0.041	0.044	0.068	0.046	0.038	0.057	**0.030**	0.034	**0.023**
VT1000	E_ξ ↑	0.921	0.933	0.925	0.944	0.947	0.935	0.937	0.948	0.944	0.945	0.935	0.945	**0.952**	**0.954**	0.949	**0.958**
	S_α ↑	0.910	0.915	0.918	0.923	0.938	0.926	0.929	0.922	0.932	0.938	0.925	0.936	**0.945**	**0.941**	0.932	**0.942**
	F_β^ω ↑	0.804	0.856	0.878	0.900	0.894	0.891	0.895	0.902	0.906	0.913	0.887	0.909	**0.921**	**0.925**	0.907	**0.927**
	MAE ↓	0.034	0.027	0.024	0.023	0.018	0.022	0.021	0.021	0.019	**0.017**	0.023	**0.017**	**0.015**	**0.015**	0.018	**0.014**
un-VT1000	E_ξ ↑	0.876	0.919	0.853	0.922	0.938	0.825	0.927	0.943	0.929	0.891	0.919	**0.940**	0.863	0.938	**0.944**	**0.954**
	S_α ↑	0.873	0.904	0.817	0.914	**0.936**	0.800	0.920	0.915	0.914	0.879	0.910	**0.932**	0.875	0.931	0.925	**0.941**
	F_β^ω ↑	0.735	0.830	0.730	0.833	0.890	0.701	0.877	0.889	0.833	0.810	0.853	**0.902**	0.758	**0.902**	**0.897**	**0.923**
	MAE ↓	0.051	0.033	0.069	0.031	**0.018**	0.077	0.025	0.023	0.028	0.045	0.028	0.020	0.042	**0.019**	0.021	**0.014**
VT821	E_ξ ↑	0.842	0.895	0.909	0.912	0.926	0.896	0.919	0.912	0.918	**0.929**	0.911	0.919	**0.929**	**0.936**	0.917	**0.932**
	S_α ↑	0.810	0.871	0.884	0.881	0.904	0.875	0.899	0.876	0.891	**0.906**	0.878	0.891	**0.912**	**0.913**	0.883	**0.906**
	F_β^ω ↑	0.627	0.760	0.821	0.829	0.818	0.801	0.841	0.823	0.835	0.849	0.809	0.835	**0.863**	**0.873**	0.817	**0.859**
	MAE ↓	0.716	0.045	0.038	0.038	0.030	0.043	0.030	0.033	0.029	0.026	0.033	0.033	**0.024**	**0.023**	0.033	**0.025**
un-VT821	E_ξ ↑	0.818	0.888	0.801	0.875	0.905	0.790	0.889	0.908	0.899	0.873	0.888	0.887	0.843	**0.910**	**0.913**	**0.929**
	S_α ↑	0.800	0.866	0.750	0.854	**0.888**	0.733	0.873	0.860	0.867	0.839	0.852	0.870	0.826	**0.894**	0.869	**0.905**
	F_β^ω ↑	0.616	0.747	0.605	0.736	**0.799**	0.575	0.788	0.799	0.741	0.736	0.746	0.795	0.670	**0.833**	0.789	**0.857**
	MAE ↓	0.073	0.048	0.089	0.063	**0.036**	0.086	0.047	0.036	0.044	0.054	0.044	**0.036**	0.056	**0.033**	0.037	**0.026**

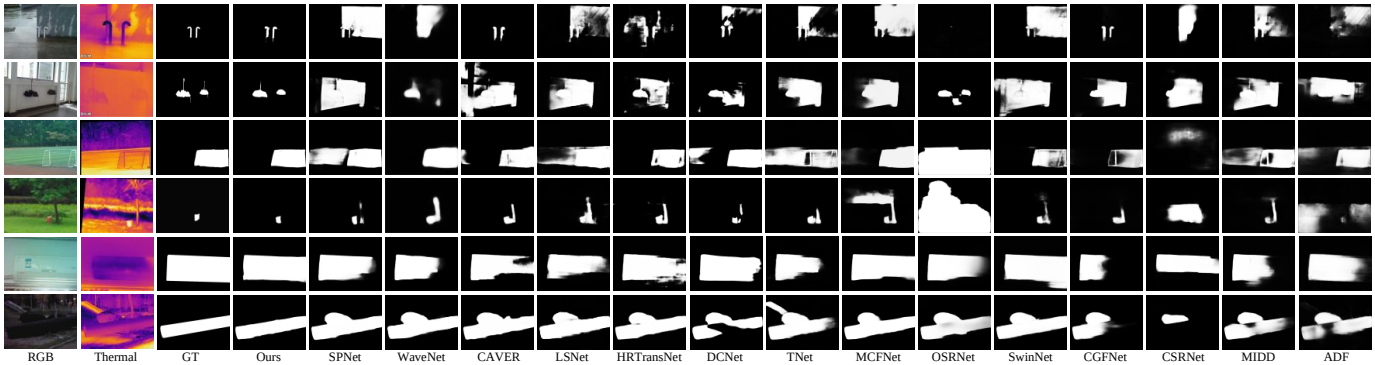


Fig. 8. Visual comparisons with other state-of-the-art methods in some challenging scenes, including bad weather (e.g., Row 1), multiple salient objects (e.g., Row 2), center bias (e.g., Row 3), small salient object (e.g., Row 4), similar foreground and background (e.g., Row 5), and low illumination (e.g., Row 6).

CGFNet [14], SwinNet [15], OSRNet [69], TNet [70], DCNet [18], MCFNet [71], HRTransNet [57], LSNet [49], CAVER [56], WaveNet [16], and SPNet [72]. For a fair comparison, we use the published code with the default parameters to implement these methods. For methods without the published source code, we directly use the results provided by the authors.

It is worth noting that our method does not need to remove the model module when handling alignment settings. There are several reasons for this. Firstly, in the case of alignment, objects in RGB and thermal modalities correspond in position and scale. Since the correlation modeling of the ACM module is performed in asymmetric window pairs of the two modalities, and the large window is obtained by expanding the small one, the corresponding information of the two modalities can still be completely covered. In addition, although the difference in window size introduces some interference information, the ACM is still able to establish sufficient multi-modal correlations in the aligned image pairs through the semantic guidance and powerful correlation operation. Secondly, being unaffected by unalignment, the AFSM module can calculate more accurate offsets to sample more relevant saliency features from the thermal modality through cascaded deformable convolutions.

1) Quantitative Evaluation. Table II shows the quantitative comparison results. It can be seen that our method overall outperforms the compared methods on the aligned, weakly aligned, and unaligned datasets. Compared with the suboptimal method (i.e., SPNet [72]), our proposed SACNet achieves average improvements of 2.5%, 1.2%, 3.2%, and 24.2% on the four evaluation metrics (i.e., E_ξ, S_α, F_β^ω, and MAE) of the seven datasets, respectively. By replacing our backbone with CNN-based ResNet50, our method still has comparable performance, especially compared with methods [56], [70], [71] that also use ResNet50 as the backbone, which further demonstrates the effectiveness of our method. In addition,

TABLE III
ABLATION STUDIES ON ONE UNALIGNED, THREE WEAKLY ALIGNED, AND THREE ALIGNED DATASETS. THE BEST RESULTS ARE MARKED WITH **BOLD**.

models	UVT2000			un-VT5000			un-VT1000			un-VT821			VT5000			VT1000			VT821		
	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$
SACNet	**0.812**	**0.640**	**0.036**	**0.949**	**0.876**	**0.023**	**0.954**	**0.923**	**0.014**	**0.929**	**0.857**	**0.026**	**0.957**	**0.888**	**0.021**	**0.958**	**0.927**	**0.014**	**0.932**	**0.859**	**0.025**
w/o ACM	0.777	0.603	0.049	0.939	0.861	0.028	0.947	0.912	0.017	0.915	0.836	0.031	0.943	0.867	0.026	0.949	0.915	0.016	0.917	0.841	0.030
w/o AWP	0.779	0.607	0.049	0.942	0.865	0.027	0.947	0.916	0.016	0.917	0.839	0.030	0.950	0.878	0.023	0.951	0.919	0.015	0.923	0.846	0.028
w/o SGM	0.790	0.612	0.044	0.944	0.869	0.026	0.948	0.914	0.016	0.918	0.844	0.029	0.951	0.881	0.023	0.953	0.919	0.015	0.923	0.848	0.028
w/o AFSM	0.785	0.618	0.045	0.941	0.865	0.027	0.948	0.913	0.016	0.919	0.840	0.030	0.946	0.873	0.024	0.949	0.914	0.016	0.922	0.845	0.029

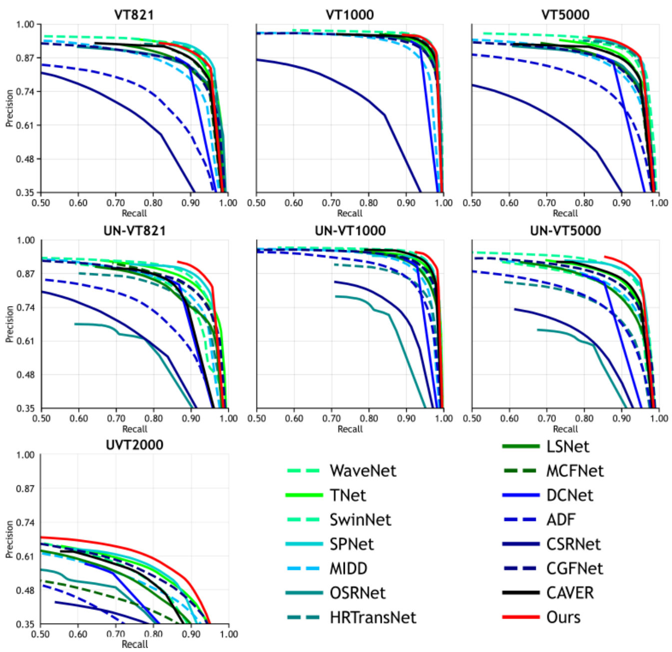


Fig. 9. The precision-recall (PR) curves of our method and 14 compared methods.

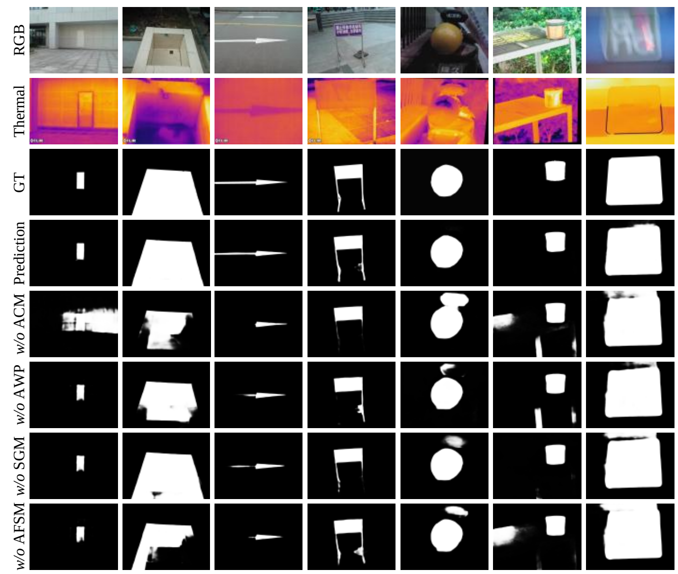


Fig. 10. Visual ablation experiments for each component in different scenes. The first to fourth columns are unaligned samples, the fifth and sixth columns are weakly aligned samples, and the seventh column are aligned samples.

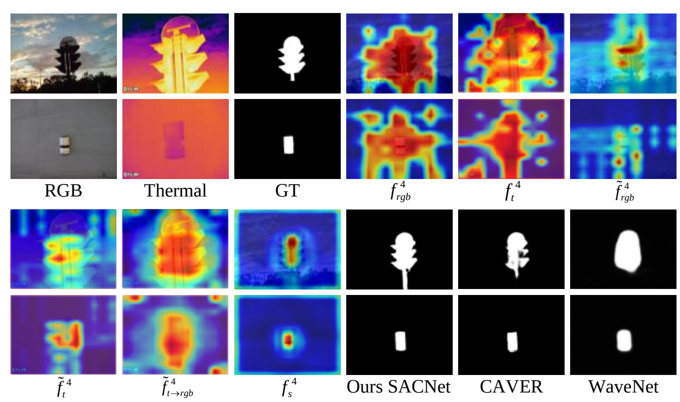


Fig. 11. Feature visualization for ACM and AFSM in the highest layer of backbone. f_s^i and f_t^4 are the extracted highest-layer features, $\tilde{f}_{rgb}^4$ and $\tilde{f}_t^4$ are features enhanced by ACM, $\tilde{f}_{t\to rgb}^4$ is the sampled thermal feature by AFSM, and f_s^4 is the integrated multi-modal feature. The comparison with advanced methods (i.e., CAVER [56] and WaveNet [16]) demonstrates the effectiveness of our method.

the computational complexity of our model is high, which mainly comes from the correlation operation and the cascaded deformable convolutions. Comparing the results on the aligned datasets and their corresponding weakly aligned datasets, we find that all compared methods suffer severe performance degradation on the weakly aligned datasets. This is mainly because they fail to establish robust correlations between the unaligned and weakly aligned modalities. In contrast, the performance of our method on the three weakly aligned datasets is closed to that on the corresponding aligned ones. This indicates that our method can better maintain the performance with a slight interference from misalignment. For the newly constructed UVT2000 dataset, the overall performance of all methods is inferior to that of the existing datasets, which indicates that alignment-free RGBT SOD is challenging and has great potential. Compared to the suboptimal method DCNet [18] on the challenging UVT2000 dataset, the minimum percentage gains on the four evaluation metrics (i.e., E_ξ, S_α, F_β^ω, and MAE) of our method are 7.8%, 10.4%, 24.8%, and 36.1%, respectively. Furthermore, the PR curves of our method and the compared methods are presented in Fig. 9. It can be seen that the curves of our method are overall more upward on the seven datasets, indicating that the saliency maps predicted by our method have a higher confidence and accuracy.

2) Qualitative Evaluation. Fig. 8 visually illustrates the qualitative comparison under various challenging scenes, including bad weather (e.g., Row 1), multiple salient objects (e.g., Row 2), center bias (e.g., Row 3), small salient object (e.g., Row 4), similar foreground and background (e.g., Row 5), and low illumination (e.g., Row 6). For a comprehensive

TABLE IV
ABLATION STUDIES ON SMALL WINDOW SIZE M AND LARGE WINDOW SIZE N OF ASYMMETRIC WINDOW PAIRS. THE BEST RESULTS ARE MARKED WITH **BOLD**.

Models	UVT2000			un-VT821			VT821		
	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$
SACNet (M=4, N=6)	**0.812**	**0.640**	**0.036**	**0.929**	**0.857**	**0.026**	**0.932**	**0.859**	**0.025**
M=4, N=4	0.789	0.621	0.040	0.920	0.845	0.029	0.928	0.852	0.028
M=6, N=6	0.789	0.613	0.044	0.920	0.844	0.029	0.928	0.853	0.026
M=3, N=4	0.798	0.625	0.039	0.923	0.850	0.028	0.928	0.853	0.026
M=2, N=4	0.801	0.629	0.038	0.923	0.841	0.029	0.930	0.841	0.028
M=2, N=6	0.773	0.601	0.050	0.914	0.837	0.031	0.919	0.842	0.031

TABLE V
ABLATION STUDIES ON THE NUMBER OF CASCADED DEFORMABLE CONVOLUTIONS N. THE BEST RESULTS ARE MARKED WITH **BOLD**.

Models	UVT2000			un-VT821			VT821		
	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$
SACNet (N=4)	**0.812**	**0.640**	**0.036**	**0.929**	**0.857**	**0.026**	**0.932**	**0.859**	**0.025**
N=1	0.793	0.621	0.043	0.921	0.847	0.029	0.923	0.849	0.027
N=2	0.795	0.621	0.042	0.924	0.850	0.030	0.926	0.852	0.027
N=3	0.794	0.625	0.040	0.926	0.853	0.029	0.929	0.854	0.026
N=5	0.802	0.631	0.040	0.927	0.856	0.028	0.929	0.857	**0.025**

comparison, the image pairs in the first two rows are from our newly constructed UVT2000 dataset, the third and fourth rows are from the weakly aligned datasets, and the last two rows are from the manually aligned datasets. It can be seen that our method is able to predict the salient regions accurately compared to both the method (i.e., DCNet [18]) for weakly aligned image pairs and the transformer-based methods (i.e., SPNet [72], WaveNet [16], CAVER [56], HRTransNet [57], and SwinNet [15]). This shows that our method can establish strong correlations between both aligned and unaligned modalities and make full use of the multi-modal complementary information.

E. Ablation Studies

In this section, we perform ablation studies to illustrate the effectiveness of the components in our method. The results for unaligned, weakly aligned, and aligned datasets are shown in Table III, in which the first line (i.e., SACNet) shows the performance of our full model. 'w/o' means to disable the corresponding component. In addition, more detailed ablation experiments for asymmetric window partition (AWP) and associated feature sampling module (AFSM) are shown in Table III and Table V, respectively. We also present some ablation analyses about the visualization results in Fig. 10 and Fig. 11.

1) Effectiveness of ACM. In order to verify the effectiveness of the proposed ACM, we directly remove it, denoted as 'w/o ACM' in Table III. This means that the correlation of the two modalities cannot be modeled adequately, especially on the unaligned and weakly aligned datasets. Compared with our full model SACNet, the performance on MAE metric drops by 36.1%, 21.7%, 21.4%, and 19.2% on UVT2000, un-VT5000, un-VT1000, and un-VT821 datasets, respectively. On the aligned VT821 dataset, the performance on the three evaluation metrics (E_ξ, F_β^ω, and MAE) separately decreases by 1.6%, 2.1%, and 20.0%. The corresponding visual results in the fifth row of Fig. 10 prove that without the correlation modeling of ACM, the model will focus on irrelevant information and introduce some noise. Fig. 11 further demonstrates the visualization of RGB and thermal features (i.e., f_{rgb}^4 and $\tilde{f}_{rgb}^4$, f_t^4 and $\tilde{f}_t^4$) before and after the ACM. With the correlation modeling of ACM, both the RGB and thermal features focus more on the objects, alleviating the interference of misalignment. This confirms that the ACM can effectively model multi-modal correlations and exploit their complementarity for salient regions.

2) Effectiveness of AWP. We also directly disable the asymmetric window pairs, which means that the correlation modeling is performed within two whole feature maps. The results in the third row (i.e., w/o AWP) of Table III and the sixth row of Fig. 10 demonstrate the positive effect of the AWP strategy. In particular, on the unaligned and weakly aligned datasets, without AWP, the MAE metrics decreased by an average of 20.8%. Note that the AWP is also valid for aligned datasets. The reason may be that under small window differences, irrelevant information is not introduced too much, and richer feature representations can be learned through the asymmetric window attention to facilitate the identification of salient regions.

The window size of asymmetric window pairs is an important hyper-parameter of the proposed asymmetric window partition (AWP). Therefore, we set five pairs of values for the size of small window M and large window N as candidate values, and record the performance of the five variants on both aligned and unaligned datasets. The results are shown in Table IV, in which the first line (i.e., SACNet) shows the performance of our complete model with a small window of size 4 (i.e., $M = 4$) and a large window of size 6 (i.e., $N = 6$). We first replace the proposed asymmetric window pairs with symmetric ones. The results are shown in the second (i.e., $M = 4$, $N = 4$) and third (i.e., $M = 6$, $N = 6$) rows of Table IV. Compared with the complete model, we can find that the asymmetric window pair works on both aligned and unaligned datasets. In addition, we also conduct experiments on the special case with a large window size difference (i.e., $M = 2$, $N = 6$), and the result is shown in the last row of Table IV. It can be seen that the large size difference between window pairs can cause performance degradation. This is mainly because the large window size difference introduces too much irrelevant information and background noise, disturbing the correlation modeling of salient regions.

3) Effectiveness of SGM. To observe the effectiveness of the SGM, we remove it and perform the correlation operation directly on the original asymmetric window pairs. The results are shown in the forth row (i.e., w/o SGM) of Table III and. By comparing with the full model, it can be found that the SGM improves the three indicators on all seven datasets, especially obtains an average gain of 14.3% on the MAE metric for the unaligned and weakly aligned datasets. The visual results in the seventh row of Fig. 10 show that without SGM, the model is able to segment rough but inaccurate salient regions. This suggests that semantic information indeed guides the correlation modeling of the ACM to focus on salient regions, thereby reducing missed and false detection.

TABLE VI
QUANTITATIVE COMPARISON OF E-MEASURE (E_ξ), S-MEASURE (S_α), WEIGHED F-MEASURE (F_β^ω), AND MEAN ABSOLUTE ERROR (MAE) ON SIX RGBD DATASETS. THE BEST THREE RESULTS ARE MARKED WITH **RED**, **GREEN**, AND **BLUE**.

	Method	CCAFNet$_{21}$ [74]	CDNet$_{21}$ [75]	HAINet$_{21}$ [76]	DFM$_{21}$ [77]	SPNet$_{21}$ [78]	RD3D$_{21}$ [79]	DSA2F$_{21}$ [80]	DCF$_{21}$ [81]	VST$_{21}$ [55]	MobileSal$_{22}$ [64]	SSL$_{22}$ [82]	DIGRNet$_{22}$ [83]	LSNet$_{23}$ [49]	CAVER$_{23}$ [56]	PICRNet$_{23}$ [84]	SACNet Ours	SACNet Ours
	Backbone	VGG16	VGG16	VGG16	MobileNet-v2	Res2Net50	I3DResNet	VGG19	ResNet50	T2T-ViT	MobileNet-v2	VGG16	ResNet50	MobileNet-v2	ResNet50	SwinT	ResNet50	SwinB
	FPS ↑	88	86	11	252	50	94	24	57	69	268	52	33	316	67	63	19	27
	Parameters(M) ↓	41.8	32.4	59.8	2.2	150.3	28.9	34.0	97.0	53.5	6.5	74.2	166.7	4.6	55.8	86.0	530.9	327.7
DUT	E_ξ ↑	0.940	0.936	0.937	0.898	0.876	0.949	0.950	0.952	**0.960**	0.936	0.927	0.948	0.927	**0.955**	**0.967**	0.953	**0.967**
	S_α ↑	0.904	0.905	0.909	0.856	0.803	**0.932**	0.921	0.924	**0.943**	0.896	0.889	0.926	0.886	0.931	**0.943**	0.923	**0.946**
	F_β^ω ↑	0.884	0.878	0.887	0.795	0.747	0.913	0.914	0.913	0.926	0.869	0.859	0.902	0.856	**0.920**	**0.935**	0.912	**0.944**
	MAE ↓	0.037	0.039	0.038	0.062	0.085	0.031	0.030	0.030	**0.025**	0.044	0.046	0.033	0.049	0.029	**0.022**	0.030	**0.021**
NJUD	E_ξ ↑	0.920	0.903	0.917	0.913	**0.931**	0.918	0.923	0.922	0.913	0.914	0.881	0.928	0.891	0.922	**0.930**	**0.930**	**0.941**
	S_α ↑	0.910	0.872	0.909	0.906	**0.925**	0.915	0.903	0.903	0.922	0.905	0.841	**0.932**	0.837	0.920	0.924	0.921	**0.935**
	F_β^ω ↑	0.883	0.828	0.882	0.868	**0.909**	0.890	0.889	0.884	0.892	0.874	0.786	**0.909**	0.775	0.903	**0.909**	**0.910**	**0.933**
	MAE ↓	0.037	0.054	0.039	0.042	**0.029**	0.037	0.039	0.038	0.035	0.040	0.065	**0.028**	0.074	0.032	0.030	**0.029**	**0.022**
NLPR	E_ξ ↑	0.951	0.951	0.951	0.945	0.957	0.957	0.950	0.956	0.953	0.950	0.954	0.955	0.955	**0.959**	**0.965**	0.951	**0.964**
	S_α ↑	0.922	0.925	0.921	0.923	0.928	0.929	0.918	0.921	**0.931**	0.920	0.919	**0.935**	0.918	0.929	**0.935**	0.920	**0.935**
	F_β^ω ↑	0.883	0.886	0.884	0.876	**0.899**	0.894	0.889	0.892	0.891	0.878	0.885	0.895	0.881	**0.899**	**0.911**	0.888	**0.917**
	MAE ↓	0.026	0.025	0.025	0.026	**0.021**	**0.022**	0.024	0.023	0.024	0.025	0.027	0.023	0.024	**0.022**	**0.019**	0.024	**0.019**
SSD	E_ξ ↑	**0.915**	0.849	0.843	0.871	**0.910**	0.905	0.904	0.898	0.907	0.898	0.833	0.889	0.902	**0.915**	**0.915**	**0.915**	**0.929**
	S_α ↑	0.876	0.798	0.769	0.814	0.871	0.863	0.876	0.852	**0.889**	0.863	0.745	0.866	0.856	0.874	**0.878**	0.876	**0.896**
	F_β^ω ↑	**0.839**	0.706	0.682	0.733	0.831	0.794	0.836	0.800	0.836	0.804	0.638	0.804	0.796	0.826	**0.837**	0.835	**0.870**
	MAE ↓	0.054	0.073	0.101	0.076	**0.044**	0.052	0.047	0.053	0.045	0.052	0.100	0.053	0.055	**0.044**	0.046	**0.042**	**0.032**
SIP	E_ξ ↑	0.915	0.913	0.924	0.919	0.930	0.919	0.908	0.920	**0.936**	0.914	0.921	0.918	0.911	0.927	0.916	**0.932**	**0.934**
	S_α ↑	0.876	0.872	0.886	0.883	0.894	0.885	0.861	0.873	**0.903**	0.873	0.880	0.885	**0.909**	0.893	0.865	0.888	**0.896**
	F_β^ω ↑	0.839	0.839	0.860	0.844	0.873	0.852	0.838	0.850	**0.878**	0.837	0.851	0.849	**0.877**	0.874	0.838	0.871	**0.889**
	MAE ↓	0.054	0.056	0.049	0.051	**0.043**	0.049	0.057	0.051	**0.040**	0.054	0.049	0.053	**0.040**	**0.043**	0.056	0.044	**0.039**
STERE	E_ξ ↑	0.921	0.929	**0.930**	0.912	**0.930**	0.926	0.928	**0.931**	0.916	0.916	0.923	0.927	0.913	**0.931**	**0.937**	**0.930**	0.929
	S_α ↑	0.891	0.907	0.909	0.898	0.907	0.911	0.897	0.905	0.913	0.903	0.897	**0.916**	0.871	0.914	**0.920**	0.902	**0.917**
	F_β^ω ↑	0.853	0.871	0.877	0.850	0.879	0.877	0.877	0.880	0.872	0.865	0.864	0.877	0.827	**0.887**	**0.898**	0.879	**0.901**
	MAE ↓	0.044	0.039	0.038	0.045	0.037	0.038	0.038	0.037	0.038	0.041	0.042	0.038	0.054	**0.034**	**0.031**	0.038	**0.030**

4) Effectiveness of AFSM. We also replace the AFSM with the concatenation and convolution operations, which means that the spatially inconsistent features of the two modalities are directly integrated. The results in the last row (i.e., w/o AFSM) of Table III and Fig. 10 demonstrate the effectiveness of AFSM. On all seven datasets, without AFSM, the three evaluation metrics (i.e., E_ξ, F_β^ω, and MAE) drop by an average of 1.4%, 1.9%, and 15.4%, respectively. As shown in Fig. 11, the thermal feature (i.e., $\tilde{f}_{t\to rgb}^4$) sampled according to the RGB feature captures the salient regions completely. Based on this, the integrated multi-modal feature (i.e., f_s^4) can locate the salient regions accurately.

We also complement the ablation experiments on the number of cascaded deformable convolutions in AFSM, with the results shown in Table V. As the number of deformable convolutions increases, AFSM can sample more relevant multi-modal features for accurate integration, with improved performance. However, when the number of deformable convolutions increases to 5, the model shows a decreasing trend. This is mainly because too many deformable convolutions introduce a large number of parameters, leading to overfitting of the model.

F. Experiment on RGBD SOD Datasets

RGBD salient object detection is another multi-modal SOD task, which utilizes the complementary information of RGB images and depth maps. To further verify the effectiveness of our method, we perform experiments on six RGBD SOD datasets, including DUT-RGBD (i.e., DUT) [61], NJUD [60], NLPR [34], SIP [62], SSD [85] and STERE [86].

1) Datasets: NJUD contains 1985 pairs of visible images and depth maps that are collected from the Internet, 3D movies, and stereo photos. DUT-RGBD contains 1200 pairs of images captured from both indoor and outdoor scenes, which include a large number of image clutter challenges. NJUD and NLPR consist of 1985 and 1000 pairs of images separately, which contain more scenarios with the challenge of scale variation. SIP incorporates 929 paired images about persons from different angles. SSD comprises 80 pairs of RGBD images from several movies. STERE collects 1000 pairs of RGBD images from the Internet. Following [80], we take a collection of 700 samples from NLPR, 1485 samples from NJUD, and 800 samples from DUT-RGBD as our RGBD training set.

2) Experiment Setup: We compare our method with 15 state-of-the-art RGBD SOD methods, including CCAFNet [74], CDNet [75], HAINet [76], DFM [77], SPNet [78], RD3D [79], DSA2F [80], DCF [81], VST [55], MobileSal [64], SSL [82], DIGRNet [83], LSNet [49], CAVER [56], and PICRNet [84]. All predicted results and code used in the experiments are released by the authors.

3) Quantitative Evaluation: Table VI shows the quantitative comparison results in terms of the four evaluation metrics (i.e., E_ξ, S_α, F_β^ω, and MAE). It can be seen that our network outperforms all compared methods on the six datasets, except for E_ξ and S_α on the SIP [62] and STERE [86] datasets, and E_ξ on the NLPR dataset. Compared with the suboptimal method (i.e., PICRNet), our method achieve average improvements of the six datasets on the four evaluation metrics (E_ξ, S_α, F_β^ω, and MAE) by 0.9%, 1.2%, 2.4%, and 22.0%, respectively. By replacing the backbone with ResNet50, our method is still comparable to the advanced method CAVER, which also uses ResNet50 as the backbone. This demonstrates that our method is also able to model the strong correlation between RGB and depth modalities and fully exploit their complementary information for saliency prediction.

G. Experiment on single-modal SOD Datasets

In order to demonstrate the advantages and applications of our method in more scenarios, we also compare it with some recent advanced single-modal SOD methods, with the results reported in Table VII. Considering that the single-modal SOD task only uses the RGB modality, we replace the original

TABLE VII
QUANTITATIVE COMPARISON OF E-MEASURE (E_ξ), S-MEASURE (S_α), WEIGHED F-MEASURE (F_β^ω), AND MEAN ABSOLUTE ERROR (MAE) ON FIVE SINGLE-MODAL DATASETS. THE BEST THREE RESULTS ARE MARKED WITH **RED**, **GREEN**, AND **BLUE**.

	Method	VST$_{21}$ [55]	ICON$_{22}$ [7]	EDN$_{22}$ [87]	MENet$_{23}$ [88]	BBRF$_{23}$ [89]	SelfReformer$_{23}$ [6]	SACNet Ours
	Backbone	T2T-ViT	SwinB	ResNet50	ResNet50	SwinB	PVT-v2	SwinB
	FPS ↑	**100**	57	**123**	-	**87**	69	27
	Parameters(M) ↓	**32.2**	92.4	**42.8**	-	74.1	**44.6**	327.7
DUTS	E_ξ ↑	0.892	**0.930**	0.908	0.921	**0.927**	0.921	**0.933**
	S_α ↑	0.896	**0.917**	0.892	0.905	0.909	**0.911**	**0.920**
	F_β^ω ↑	0.828	**0.886**	0.845	0.870	**0.886**	**0.872**	**0.899**
	MAE ↓	0.037	**0.025**	0.035	0.028	**0.025**	**0.027**	**0.023**
ECSSD	E_ξ ↑	0.918	**0.932**	0.929	0.925	**0.934**	0.929	**0.937**
	S_α ↑	0.932	**0.941**	0.927	0.928	**0.939**	0.936	**0.948**
	F_β^ω ↑	0.910	**0.936**	0.918	0.920	**0.944**	0.926	**0.951**
	MAE ↓	0.033	**0.023**	0.032	0.031	**0.022**	0.027	**0.019**
OMRON	E_ξ ↑	0.861	**0.898**	0.879	0.882	**0.891**	0.889	**0.904**
	S_α ↑	0.850	**0.869**	0.849	0.850	**0.861**	**0.861**	**0.871**
	F_β^ω ↑	0.755	**0.804**	0.770	0.771	**0.803**	0.784	**0.814**
	MAE ↓	0.050	**0.043**	0.049	0.045	**0.044**	**0.043**	**0.041**
HKU-IS	E_ξ ↑	0.953	**0.965**	0.956	**0.960**	**0.965**	0.959	**0.969**
	S_α ↑	0.928	**0.935**	0.924	0.927	**0.932**	0.931	**0.939**
	F_β^ω ↑	0.897	**0.925**	0.908	0.917	**0.932**	0.915	**0.938**
	MAE ↓	0.029	**0.022**	0.026	0.023	**0.020**	0.024	**0.018**
PASCAL-S	E_ξ ↑	0.843	**0.875**	0.870	0.870	0.873	**0.879**	**0.880**
	S_α ↑	0.871	**0.885**	0.864	0.871	0.878	**0.881**	**0.883**
	F_β^ω ↑	0.822	**0.860**	0.833	0.844	**0.862**	0.854	**0.866**
	MAE ↓	0.061	**0.048**	0.061	0.054	**0.049**	0.051	**0.047**

RGBT image pair with two identical RGB images as input to our network.

1) Datasets: We evaluate our method on five representative single-modal SOD datasets, including DUTS [90] (10,553 training images and 5,019 testing images), OMRON [91] (5,168 images), ECSSD [92] (1,000 images), HKU-IS [93] (4,447 images), and PASCAL-S [94] (850 images). Following [6], [7], we use the training set of DUTS to train our single-modal SOD model.

2) Experiment Setup: We compare our method with 6 state-of-the-art single-modal SOD methods, including VST [55], ICON [7], END [87], MENet [88], BBRF [89], and SelfReformer [6]. All predicted results and code used in the experiments are released by the authors.

3) Quantitative Evaluation: The results in Table VII show that our method outperforms all compared methods on the five datasets, except for the S-measure metric on the PASCAL-S dataset. For example, compared with the second best method (i.e., ICON), our method achieves an average improvement of 0.5%, 0.3%, 1.3%, and 11.8% for the four evaluation metrics (i.e., E_ξ, S_α, F_β^ω, and MAE) across the five datasets. This demonstrates that our method is also applicable to the single-modal SOD task.

H. Failure Cases and Future Work

Although our method achieves superior performance on unaligned image pairs through modeling robust multi-modal correlations, it still fails in some extreme challenging scenarios. Fig. 12 shows the visual results of our method and other advanced methods (i.e., SPNet [72], WaveNet [16], and SwinNet [15]) for some typical failure cases. For the examples in the first and second rows, the salient objects are hollow and their foreground and background regions are intertwined, which makes it difficult for our method and other methods to make refined predictions. The salient objects in the third and forth rows are challenged by strong light, which interferes

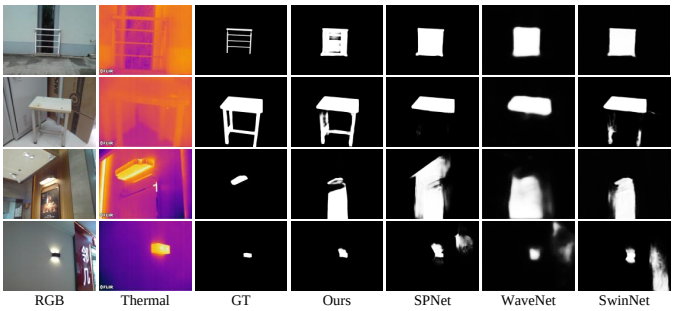


Fig. 12. Visual results of our SACNet and other advanced methods (i.e., SPNet [72], WaveNet [16], and SwinNet [15]) in some typical failure cases, including hollow objects (i.e., Rows 1 and 2), strong light interference (i.e., Rows 3 and 4).

with our correlation modeling of salient regions, resulting in inaccurate localization and segmentation of the salient objects. These failure cases reveal the limitations of our method, which still struggles to deal with the challenging scenarios in unaligned image pairs. To this end, in future work, we will design fusion modules to handle different challenges. By integrating the feature representations of these specific fusion modules into the correlation modeling, the robustness of our method in challenging scenarios can be further improved.

In addition, UVT2000 serves as the first unaligned RGBT SOD dataset for research on alignment-free RGBT SOD, it still has some limitations. We will further expand the UVT2000 dataset in future work to better reflect the diversity and complexity in real-world applications. Specifically, in terms of scale, we will capture larger-scale unaligned RGBT image pairs that far exceed the size of any existing multi-modal SOD datasets to improve the diversity of the UVT2000 dataset. In terms of scenes and circumstances, we will cover more practical scenes, such as traffic scenes, drone scenes, and lake scenes, to promote extensive research on unaligned RGBT image pairs for real-world applications. In terms of challenges, we will capture and annotate more representative challenging scenarios that are specific to unaligned RGBT image pairs to enhance the complexity of the UVT2000 dataset. In this way, the UVT2000 dataset will be improved as a more solid foundation for comprehensive research on alignment-free RGBT SOD.

VI. CONCLUSION

In this paper, we explore the saliency complementarity in unaligned visible-thermal image pairs and propose a semantics-guided asymmetric correlation network, which models robust correlations between RGB and thermal modalities without manual alignment. To this end, two components (i.e., ACM and AFSM) are proposed. The ACM is able to establish comprehensive multi-modal correlations specific to salient regions, and the AFSM can sample relevant thermal features conditioned on corresponding RGB features for accurate integration. Additionally, we contribute a novel unaligned RGBT SOD benchmark dataset called UVT2000, which provides a challenging platform and facilitates research on alignment-free RGBT SOD. Experimental results demonstrate that our method achieves state-of-the-art performance on

both aligned and unaligned datasets. The overall performance of existing methods and our method on the newly constructed UVT2000 dataset shows the great potential of alignment-free RGBT SOD.

REFERENCES

[1] L. Shao and M. Brady, "Specific object retrieval based on salient regions," *Pattern Recognit.*, vol. 39, no. 10, pp. 1932–1948, 2006.

[2] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3586–3593.

[3] P. Zhang, W. Liu, D. Wang, Y. Lei, H. Wang, and H. Lu, "Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps," *Pattern Recognit.*, vol. 100, p. 107130, 2020.

[4] Y. Kong, Y. Wang, A. Li, and Q. Huang, "Self-sufficient feature enhancing networks for video salient object detection," *IEEE Transactions on Multimedia*, vol. 25, pp. 557–571, 2023.

[5] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8554–8564.

[6] Y. K. Yun and W. Lin, "Towards a complete and detail-preserved salient object detection," *IEEE Transactions on Multimedia*, pp. 1–15, 2023.

[7] M. Zhuge, D.-P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao, "Salient object detection via integrity learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3738–3752, 2022.

[8] Y. Piao, W. Wu, M. Zhang, Y. Jiang, and H. Lu, "Noise-sensitive adversarial learning for weakly supervised salient object detection," *IEEE Transactions on Multimedia*, vol. 25, pp. 2888–2897, 2022.

[9] D. Zhu, R. Hu, S. Song, X. Guo, X. Li, and Z. Wang, "Cross-illumination video anomaly detection benchmark," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 2516–2525.

[10] X. Xu, S. Wang, Z. Wang, X. Zhang, and R. Hu, "Exploring image enhancement for salient object detection in low light images," *ACM transactions on multimedia computing, communications, and applications (TOMM)*, vol. 17, no. 1s, pp. 1–19, 2021.

[11] B. Wan, X. Zhou, Y. Sun, T. Wang, C. Lv, S. Wang, H. Yin, and C. Yan, "Mffnet: Multi-modal feature fusion network for vdt salient object detection," *IEEE Transactions on Multimedia*, pp. 1–14, 2023.

[12] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGB-T salient object detection via fusing multi-level CNN features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, 2020.

[13] Z. Tu, Z. Li, C. Li, Y. Lang, and J. Tang, "Multi-interactive dual-decoder for rgb-thermal salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 5678–5691, 2021.

[14] J. Wang, K. Song, Y. Bao, L. Huang, and Y. Yan, "Cgfnet: Cross-guided fusion network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2949–2961, 2022.

[15] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "Swinnet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4486–4497, 2022.

[16] W. Zhou, F. Sun, Q. Jiang, R. Cong, and J. Hwang, "Wavenet: Wavelet network with knowledge distillation for RGB-T salient object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 3027–3039, 2023.

[17] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "Rgb-t image saliency detection via collaborative graph learning," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 160–173, 2019.

[18] Z. Tu, Z. Li, C. Li, and J. Tang, "Weakly alignment-free RGBT salient object detection with deep correlation network," *IEEE Trans. Image Process.*, vol. 31, pp. 3752–3764, 2022.

[19] G. Wang, C. Li, Y. Ma, A. Zheng, J. Tang, and B. Luo, "Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach," in *Image and Graphics Technologies and Applications: 13th Conference on Image and Graphics Technologies and Applications, IGTA 2018, Beijing, China, April 8–10, 2018, Revised Selected Papers 13*. Springer, 2018, pp. 359–369.

[20] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "RGBT salient object detection: A large-scale dataset and benchmark," *CoRR*, vol. abs/2007.03262, 2020.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] L. Liu, C. Li, A. Zheng, J. Tang, and Y. Xiang, "Non-aligned rgbt tracking via joint temporal-iterated homography estimation and multimodal transformer fusion," in *2023 10th international conference on computational science/intelligence and applied informatics (CSII)*, 2023.

[23] Z. Zhao, Y. Zhang, C. Li, Y. Xiao, and J. Tang, "Thermal uav image super-resolution guided by multiple visible cues," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[24] D. Jiang and M. Ye, "Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2787–2797.

[25] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3239–3259, 2022.

[26] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3907–3916.

[27] J. Liu, Q. Hou, Z. Liu, and M. Cheng, "Poolnet+: Exploring the potential of pooling for salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 887–904, 2023.

[28] H. Xiao, J. Feng, Y. Wei, M. Zhang, and S. Yan, "Deep salient object detection with dense connections and distraction diagnosis," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3239–3251, 2018.

[29] N. Liu, J. Han, and M. Yang, "Picanet: Pixel-wise contextual attention learning for accurate saliency detection," *IEEE Trans. Image Process.*, vol. 29, pp. 6438–6451, 2020.

[30] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7479–7489.

[31] Z. Yao and L. Wang, "Boundary information progressive guidance network for salient object detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 4236–4249, 2021.

[32] X. Tian, J. Zhang, M. Xiang, and Y. Dai, "Modeling the distributional uncertainty for salient object detection models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 660–19 670.

[33] Y. Wang, W. Zhang, L. Wang, T. Liu, and H. Lu, "Multi-source uncertainty mining for deep unsupervised saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 727–11 736.

[34] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: A benchmark and algorithms," in *European conference on computer vision*. Springer, 2014, pp. 92–109.

[35] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "RGBD salient object detection via deep fusion," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2274–2285, 2017.

[36] T. Zhou, D. Fan, M. Cheng, J. Shen, and L. Shao, "RGB-D salient object detection: A survey," *Comput. Vis. Media*, vol. 7, no. 1, pp. 37–69, 2021.

[37] H. Song, Z. Liu, H. Du, G. Sun, O. L. Meur, and T. Ren, "Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4204–4216, 2017.

[38] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network," in *European conference on computer vision*. Springer, 2020, pp. 275–292.

[39] R. Cong, Q. Lin, C. Zhang, C. Li, X. Cao, Q. Huang, and Y. Zhao, "Cir-net: Cross-modality interaction and refinement for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 6800–6815, 2022.

[40] H. Wen, K. Song, L. Huang, H. Wang, and Y. Yan, "Cross-modality salient object detection network with universality and anti-interference," *Knowledge-Based Systems*, vol. 264, p. 110322, 2023.

[41] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "Cnns-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3171–3183, 2018.

[42] M. Zhang, S. Yao, B. Hu, Y. Piao, and W. Ji, "C^2dfnet: Criss-cross dynamic filter network for rgb-d salient object detection," *IEEE Transactions on Multimedia*, vol. 25, pp. 5142–5154, 2022.

[43] L. Huang, K. Song, J. Wang, M. Niu, and Y. Yan, "Multi-graph fusion and learning for RGBT image saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1366–1377, 2022.

[44] G. Chen, F. Shao, X. Chai, H. Chen, Q. Jiang, X. Meng, and Y. Ho, "Cgmdrnet: Cross-guided modality difference reduction network for

RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6308–6323, 2022.

[45] Z. Xie, F. Shao, G. Chen, H. Chen, Q. Jiang, X. Meng, and Y. Ho, "Cross-modality double bidirectional interaction and fusion network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4149–4163, 2023.

[46] W. Zhou, Q. Guo, J. Lei, L. Yu, and J. Hwang, "Ecffnet: Effective and consistent feature fusion network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1224–1235, 2022.

[47] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, "Apnet: Adversarial learning assistance and perceived importance fusion network for all-day RGB-T salient object detection," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 6, no. 4, pp. 957–968, 2022.

[48] G. Liao, W. Gao, G. Li, J. Wang, and S. Kwong, "Cross-collaborative fusion-encoder network for robust rgb-thermal salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7646–7661, 2022.

[49] W. Zhou, Y. Zhu, J. Lei, R. Yang, and L. Yu, "Lsnet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images," *IEEE Trans. Image Process.*, vol. 32, pp. 1329–1340, 2023.

[50] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[51] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8126–8135.

[52] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[53] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[54] Y. Zhou, Z. Li, C. Guo, S. Bai, M. Cheng, and Q. Hou, "Srformer: Permuted self-attention for single image super-resolution," *CoRR*, vol. abs/2303.09735, 2023.

[55] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4722–4732.

[56] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "CAVER: cross-modal view-mixed transformer for bi-modal salient object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 892–904, 2023.

[57] B. Tang, Z. Liu, Y. Tan, and Q. He, "Hrtransnet: Hrformer-driven two-modality salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 728–742, 2023.

[58] S. Mohammadi, M. Noori, A. Bahri, S. G. Majelan, and M. Havaei, "Cagnet: Content-aware guidance for salient object detection," *Pattern Recognit.*, vol. 103, p. 107303, 2020.

[59] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[60] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *2014 IEEE international conference on image processing (ICIP)*. IEEE, 2014, pp. 1115–1119.

[61] W. Ji, G. Yan, J. Li, Y. Piao, S. Yao, M. Zhang, L. Cheng, and H. Lu, "DMRA: depth-induced multi-scale recurrent attention network for RGB-D saliency detection," *IEEE Trans. Image Process.*, vol. 31, pp. 2321–2336, 2022.

[62] D. Fan, Z. Lin, Z. Zhang, M. Zhu, and M. Cheng, "Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 5, pp. 2075–2089, 2021.

[63] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[64] Y. Wu, Y. Liu, J. Xu, J. Bian, Y. Gu, and M. Cheng, "Mobilesal: Extremely efficient RGB-D salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 10 261–10 269, 2022.

[65] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 270–279.

[66] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[67] N. Liu, N. Zhang, L. Shao, and J. Han, "Learning selective mutual attention and contrast for rgb-d saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9026–9042, 2021.

[68] F. Huo, X. Zhu, L. Zhang, Q. Liu, and Y. Shu, "Efficient context-guided stacked refinement network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3111–3124, 2022.

[69] F. Huo, X. Zhu, Q. Zhang, Z. Liu, and W. Yu, "Real-time one-stream semantic-guided refinement network for rgb-thermal salient object detection," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[70] R. Cong, K. Zhang, C. Zhang, F. Zheng, Y. Zhao, Q. Huang, and S. Kwong, "Does thermal really always matter for rgb-t salient object detection?" *IEEE Transactions on Multimedia*, vol. 25, pp. 6971–6982, 2022.

[71] S. Ma, K. Song, H. Dong, H. Tian, and Y. Yan, "Modal complementary fusion network for RGB-T salient object detection," *Appl. Intell.*, vol. 53, no. 8, pp. 9038–9055, 2023.

[72] Z. Zhang, J. Wang, and Y. Han, "Saliency prototype for rgb-d and rgb-t salient object detection," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3696–3705.

[73] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[74] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, "Ccafnet: Crossflow and cross-scale adaptive fusion network for detecting salient objects in rgb-d images," *IEEE Transactions on Multimedia*, vol. 24, pp. 2192–2204, 2021.

[75] W. Jin, J. Xu, Q. Han, Y. Zhang, and M. Cheng, "Cdnet: Complementary depth network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3376–3390, 2021.

[76] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3528–3542, 2021.

[77] W. Zhang, G.-P. Ji, Z. Wang, K. Fu, and Q. Zhao, "Depth quality-inspired feature manipulation for efficient rgb-d salient object detection," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 731–740.

[78] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, and L. Shao, "Specificity-preserving rgb-d saliency detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4681–4691.

[79] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, "Rgb-d salient object detection via 3d convolutional neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 2, 2021, pp. 1063–1071.

[80] P. Sun, W. Zhang, H. Wang, S. Li, and X. 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*, 2021, pp. 1407–1417.

[81] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu *et al.*, "Calibrated rgb-d salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9471–9481.

[82] X. Zhao, Y. Pang, L. Zhang, H. Lu, and X. Ruan, "Self-supervised pretraining for rgb-d salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 3463–3471.

[83] X. Cheng, X. Zheng, J. Pei, H. Tang, Z. Lyu, and C. Chen, "Depth-induced gap-reducing network for rgb-d salient object detection: an interaction, guidance and refinement approach," *IEEE Transactions on Multimedia*, vol. 25, pp. 4253–4266, 2022.

[84] R. Cong, H. Liu, C. Zhang, W. Zhang, F. Zheng, R. Song, and S. Kwong, "Point-aware interaction and cnn-induced refinement network for rgb-d salient object detection," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 406–416.

[85] C. Zhu and G. Li, "A three-pathway psychobiological framework of salient object detection using stereoscopic technology," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 3008–3014.

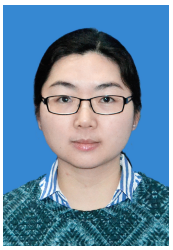
[86] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2012, pp. 454–461.
[87] Y.-H. Wu, Y. Liu, L. Zhang, M.-M. Cheng, and B. Ren, "Edn: Salient object detection via extremely-downsampled network," *IEEE Transactions on Image Processing*, vol. 31, pp. 3125–3136, 2022.
[88] Y. Wang, R. Wang, X. Fan, T. Wang, and X. He, "Pixels, regions, and objects: Multiple enhancement for salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 031–10 040.
[89] M. Ma, C. Xia, C. Xie, X. Chen, and J. Li, "Boosting broader receptive fields for salient object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1026–1038, 2023.
[90] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3796–3805.
[91] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang, "Saliency detection via graph-based manifold ranking," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3166–3173.
[92] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1155–1162.
[93] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5455–5463.
[94] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280–287.

Kunpeng Wang is currently a Ph.D. student at the School of Computer Science and Technology, Anhui University, China. He received his B.Eng. degree from the School of Software Engineering, Jinling Institute of Technology, China, in 2020. His research interests lie in computer vision and deep learning.

Danying Lin is currently a M.S. student at the College of Computer Science and Technology, Anhui University. She re ceived the B.S. degree in Internet of Things Engineering from Jiaying College in 2022. Her research interests are computer vision and deep learning.

Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively.

From 2014 to 2015, he was a Visiting Student with the School of Artificial Intelligence, Sun Yat-sen University, Guangzhou, China. He was a Post-Doctoral Research Fellow with the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. He is currently a Professor and the Ph.D. Supervisor with the School of Artificial Intelligence, Anhui University. His research interests include computer vision and deep learning.



Zhengzheng Tu received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2007 and 2015, respectively. She is currently an Associate Professor in the School of Computer Science and Technology, Anhui University. Her research interests include computer vision and deep learning.

Bin Luo received the B.Eng. degree in electronics and the M.Eng. degree in computer science from Anhui University, Hefei, China, in 1984 and 1991, respectively, and the Ph.D. degree in computer science from the University of York, York, U.K., in 2002. He is currently a full Professor with Anhui University, China. His research interests include, pattern recognition, digital image processing and cognitive computation. He chairs the IEEE Hefei Subsection. He serves as the AEiC of the journal Visual Intelligence, the EiC of the Journal of Anhui University (Natural Science Edition), the associate editor of several international journals, including Pattern Recognition, Pattern Recognition Letters, Cognitive Computation.