HRTransNet: HRFormer-Driven Two-Modality Salient Object Detection

Bin Tang, Zhengyi Liu*, Yacheng Tan, and Qian He

***Abstract*—The High-Resolution Transformer (HRFormer) can maintain high-resolution representation and share global receptive fields. It is friendly towards salient object detection (SOD) in which the input and output have the same resolution. However, two critical problems need to be solved for two-modality SOD. One problem is two-modality fusion. The other problem is the HRFormer output's fusion. To address the first problem, a supplementary modality is injected into the primary modality by using global optimization and an attention mechanism to select and purify the modality at the input level. To solve the second problem, a dual-direction short connection fusion module is used to optimize the output features of HRFormer, thereby enhancing the detailed representation of objects at the output level. The proposed model, named HRTransNet, first introduces an auxiliary stream for feature extraction of supplementary modality. Then, features are injected into the primary modality at the beginning of each multi-resolution branch. Next, HRFormer is applied to achieve forwarding propagation. Finally, all the output features with different resolutions are aggregated by intra-feature and inter-feature interactive transformers. Application of the proposed model results in impressive improvement for driving two-modality SOD tasks, e.g., RGB-D, RGB-T, and light field SOD.https://github.com/liuzywen/HRTransNet**

***Index Terms*—HRFormer, salient object detection, cross modality, RGB-D, RGB-T, light field**

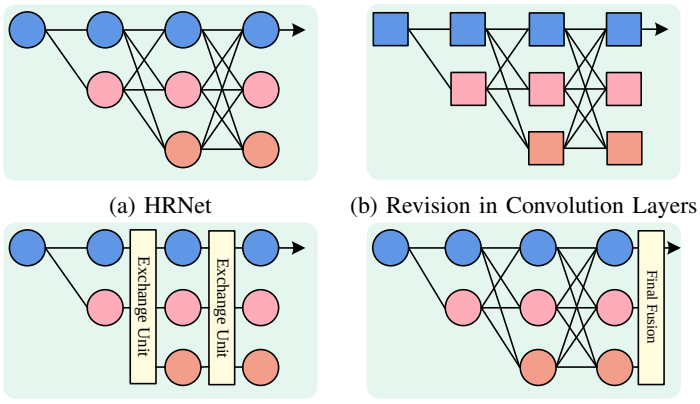
I. Introduction

Salient object detection (SOD) is an important computer vision task that automatically identifies the most attractive and conspicuous objects in a scene. SOD plays a fundamental role in image segmentation [1], [2], tracking [3], [4], cropping [5], [6], retargeting [7], activity prediction [8], etc.

It continues to be challenging to apply SOD in the primary modality "RGB image" under some conditions (e.g., poor illumination, complex backgrounds, and indistinguishable objects). RGB-Depth (RGB-D) SOD [9], RGB-Thermal (RGB-T) SOD, light field (LF) SOD [10] have gradually become cutting-edge research area because of the widespread use of depth sensors, infrared cameras, and Lytro cameras. Additional information can be obtained by using supplementary modalities: "depth image" can provide additional spatial structural information; "thermal image" can capture heat information of objects; and "focal stack" provides several focal slices at different depths. Effective fusion of color and supplementary modalities is critical for saliency detection. However, most studies to date have focused on a specific type of two-modality SOD. We explore a two-modality SOD model that is effective for RGB-D, RGB-T, and LF SOD.

The convolutional neural network (CNN) serves as a critical backbone feature extractor for a long time. The CNN takes as input an image and applies a stack of convolution and pooling layers to expand the receptive field and obtain a global view in deep layers. The CNN is friendly to image classification tasks, but not dense prediction tasks, which require a decoder to gradually recover the resolution of features.

HRNet [11] (Fig. 1 (a)) maintains high-resolution representation throughout the network. However, the expressivity of HRNet is limited by small receptive fields and strong inductive bias from cascaded convolution operations [12]. Therefore, optimized HRNets have been developed with improved stacked convolution layers (Fig. 1 (b)), the exchange units (Fig. 1 (c)), and final multi-resolution fusion (Fig. 1 (d)). The recently proposed HRFormer [13] replaces convolution layer (shown as a round unit) with transformer blocks (shown as a rectangle unit) shown in Fig. 1 (b) to improve performance. HRFormer can effectively perform dense prediction tasks by injecting global ability when maintaining local features.



(a) HRNet (b) Revision in Convolution Layers

(c) Revision in Exchange Units (d) Revision in Final Fusion

Fig. 1. HRNet (a) is improved in stacked convolution layers (b), exchange units (c), and final multi-resolution fusion (d).

This work is supported by the Natural Science Foundation of Anhui Province (1908085MF182), the Science Research Project for Graduate Student of Anhui Provincial Education Department (YJS20210047), and the Talent Research Fund Project of Hefei University(21-22RC14)(Corresponding author: Zhengyi Liu).

Bin Tang is with School of Artificial Intelligence and Big Data, Hefei University, Hefei, China(e-mail: 424539820@qq.com).

Zhengyi Liu, Yacheng Tan and Qian He are with Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei, China(e-mail: liuzywen@ahu.edu.cn,1084043983@qq.com,1819469871@qq.com).

0000–0000/00$00.00 © 2022 IEEE

In this study, we use HRFormer as a backbone network to drive two-modality SOD tasks, in particular, RGB-D, RGB-T, and LF SOD. Two issues need to be addressed: (1) How can two-modality fusion be achieved using HRFormer? (2) Can HRFormer be further improved?

Ren et al. [14] propose dual-stream HRNet to perform two-modality fusion, that is, two types of heterogeneous data (synthetic aperture radar and optical image) are combined. Features with different modalities are fused in the high-resolution branch. This design generates an insufficient receptive field. AVT [15] exploits auditory information to aid visual models in crowd counting tasks. Audio embedding is only integrated into image features in the last three-branch exchange unit. Audio modality is shallow-modeled, which is not suitable for supplementary image modalities, such as the depth image, thermal image or focal stack.

Unlike existing fusion methods, an auxiliary stream is used in this study to encode a supplementary modality and inject the modality into HRFormer at the beginning of each multi-resolution branch. The supplementary modality is deep-modeled to exploit abundant information and enhance representation.

Fig. 2 shows the final multi-resolution fusion: HRNetV1 [11] only outputs a high-resolution feature map for pose estimation; HRNetV2 [16], [17] concatenates all the output features for semantic segmentation and facial-landmark detection; HRNetV2p-V1 and HRNetV2p-V2 [16], [17] output feature pyramid representation from high-to-low or from low-to-high for object detection and classification tasks.

Unlike the aforementioned networks, a dual-direction short connection fusion module is designed in this study, that consists of four **I**ntra-feature and **I**nter-feature **I**nteractive **T**ransformers (TripleITs). The proposed module is used to optimize each resolution output feature by itself and the other resolution features, to accurately depict the details of an object.

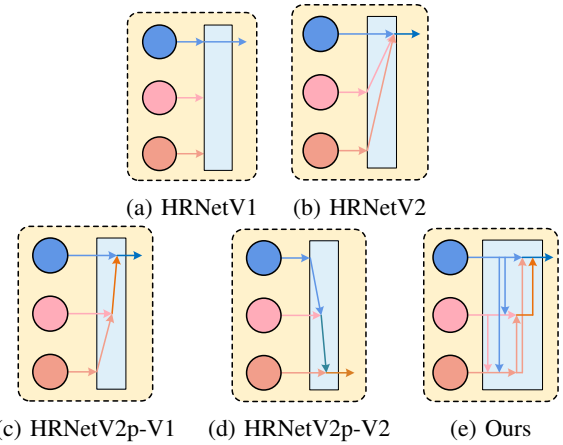


Fig. 2. Comparison of final multi-resolution fusion methods among HRNetV1, HRNetV2, HRNetV2p-V1, HRNetV2p-V2, and ours.

The four main contributions of this study are summarized below:

- A unified two-modality SOD model (HRTransNet) is proposed for RGB-D, RGB-T, and LF SOD tasks. This model can be further applied to other dense prediction tasks with two modalities. HRTransNet is built upon the HRFormer backbone and maintains high-resolution representation with a large receptive field achieved by transform blocks.
- A module is proposed to inject a supplementary modality into the primary modality by formulating the weight of each supplementary modality from a global perspective and emphasizing the coordinate-wise spatial position of the supplementary modality, thereby achieving two-modality fusion at the input level.
- A novel designed dual-direction short connection fusion module is used to optimize a resolution feature by itself and the other resolution features, thereby improving the final multi-resolution fusion of HRFormer at the output level.
- The proposed model exhibits excellent performance for the RGB-D, RGB-T, and LF datasets, which demonstrates the superiority of the model for two-modality complementary tasks.

II. Related Works

A. High-resolution network

HRNet [11] maintains high-resolution representations in the forwarding propagation process by generating feature maps with different resolutions in parallel and repeatedly conducting multi-scale fusions in the exchange unit, which is friendly to dense prediction tasks. HRNet has been widely applied for human-pose estimation [11], [18], semantic segmentation [19], facial-landmark detection [16], surface-defect detection [20], video tracking [21], image inpainting [22], remote-sensing pansharpening [23], and gaze estimation [24].

A few methods have recently been developed to optimize HRNet. Improvements are made to the stacked convolution layers, exchange unit and final multi-resolution fusion, shown in Fig. 1.

The following modifications to stacked convolution layers have been proposed: in Lite-HRNet [25], a conditional channel weight block replaces the convolution layer. Zhang et al. [18] add an attention module before the convolution layers of the last stage. Wang et al. [22] use four resolutions at the beginning rather than adding resolution gradually, and apply a shared attention map to all the resolution features of the last stage. In HRFormer [13], HRViT [12], and HR-NAS [26] the convolution layer is replaced with a transformer block [27], [28] to expand the receptive field.

The following modifications to the exchange unit have been proposed: Yang et al. [29] use an attentive multi-scale fusion module to modify feature fusion in the exchange units. This scheme is flexible for different types of information and lets the network focus on more discriminate features. Zhao et al. [30] replace the original exchange unit with a gated multi-scale fusion module to eliminate noise ambiguities. In Lite-HRNet [25] and HRViT [12], normal convolutions are replaced with depthwise separable convolutions [31] in the exchange unit. In AVT [15], a transformer-inspired attention mechanism is deployed to perform inter-branch fusion.

Fig. 2 shows the modifications proposed for final multi-resolution fusion: HRNetV1 [11] only outputs the high-resolution feature map. HRNetV2 [16], [17] concatenates all the output features. HRNetV2p [16], [17] outputs the feature pyramid representation from high-to-low or from low-to-high. Ding et al. [32] introduce a soft conditional gate module [33] at the last stage to generate dynamic weights reflecting the importance of different features for fusion. Wu et al. [19] use mixed dilated convolution to enhance the scale-awareness ability of output features and data-dependent upsampling [34] to progressively fuse multi-level features. Yang et al. [29] gradually combine low-to-high resolution features by interpolation, deconvolution, and a series of convolutions to extract the local and global features involved at all resolutions. Yang et al. [35] aggregate the feature maps of different layers using spatial attention and channel attention.

In this study, we drive the model using HRFormer in which the stacked convolution layers have been modified. Furthermore, we introduce a transformer-based fusion strategy for final multi-resolution fusion. This strategy effectively exploits the correlation between intra- and inter-features.

In addition, in the multi-modality fusion area, Ren et al. [14] propose a dual-stream HRNet to combine two types of heterogeneous data. Features with different modalities are fused in high-resolution branches. An insufficient receptive field is generated. AVT [15] embeds the audio modality into the image modality only in the last three-branch exchange unit.

Unlike existing fusion methods, in this study, a supplementary modality encoded by the auxiliary stream is injected into HRFormer at the beginning of each multi-resolution branch.

B. Two-modality salient object detection

Initially, we detect the salient objects in a color image using conventional methods [36]–[38] or CNN-based [39]–[44] methods. However, some objects are very difficult to detect, irrespective of the method used. This result may be obtained because of complex backgrounds or indistinguishable objects in a scene. Additional information that is not well represented in a color image can be obtained using supplementary modalities, e.g., the depth image (D), thermal image (T), and focal stack (FS), using widely used depth sensors, infrared imaging devices, and Lytro cameras. For example, the depth image is effective for obtaining the spatial information, the thermal image can capture the heat radiated from objects under poor illumination, and the focal stack provides focus cues at different depths. Therefore, the depth and thermal image and the focal stacks serve as supplementary modalities for improving SOD.

RGB-D, RGB-T and LF (RGB-FS) SOD all use two modality fusion to comprehensively segment objects. Early fusion [45], middle fusion [46]–[49] and late fusion [50]–[52] are three classic frameworks. The performance of RGB-D SOD can be improved using an attention mechanism [53], edge guidance [54], [55], depth calibration [56], depth estimation [57], [58], depth quality assessments [59], [60], 3D convolution [61], deformable convolution [62], automatic architecture search [63], uncertainty distribution [64], and bi-directional guidance [65], [66]. To reduce the influence of poor depth images, Cong et al. [67] use salient seed diffusion with a depth constraint and introduce a depth confidence weight [68]. Chen et al. [69] learn the potential confidence of depth map that will be aggregated into multi-modality fusion.

RGB-D SOD is studied prevalently, whereas, RGB-T and LF SOD are less exploited. The support vector machine [70], ranking algorithm [71]–[73], and graph learning [74] were initially used for RGB-T SOD. Later, CNN based methods (e.g., ECFFNet [75], MIDD [76], MMNet [77], CGFNet [78], CSRNet [79], CGMDRNet [80], and SwinNet [27]) were used to exploit two-modality fusion within attention, boundary, local and global contexts, depth-guidance, and cross-guidance perspectives. The focal stack indicates different focused regions in different depth layers for LF SOD. Early methods [81] and [82] introduced prior knowledge and weighted sparse coding. Later, ConvLSTM was employed [83]–[86] to process the focal stack due to sequential attribution. Recently widely used techniques (e.g., 3D convolution [87] and graph networks [88]) have been introduced to model information fusion within the focal stack.

The summary on RGB-D, RGB-T, and LF SOD methods, presented above shows that each of these methods can be used to perform one or two tasks [27], [77]. Few models have been designed to perform two-modality SOD. The three tasks exhibit common characteristics. A supplementary modality needs to be purified in combination with primary modality because of the presence of noise. Irrespective of the supplementary modality is used, SOD remains the dense prediction task that must to be performed to maintain the resolution between the input and output. Thus, we adapt HRFormer to a two-modality SOD task.

We also identify other two-modality SOD tasks (e.g., image-video optical flow SOD [89], text-image SOD [90], audio-image SOD [91]) and the similar but different image pair co-saliency task [92]–[94]. Therefore, two-modality SOD is worthy of further study.

C. Transformer-based salient object detection

In the past year, transformers have been successfully employed to execute computer vision tasks. GLSTR [95] is the first case in which a pure transformer-based encoder has been applied for SOD. Transformers were subsequently gradually applied in research on RGB-D SOD [9] and co-saliency [96]. RGB-D SOD has been achieved using TriTransNet [97], where three transformers with shared weights are employed to enhance the representation ability of high-layer features. TransCMD [98] decodes multi-scale and multi-modal features using a transformer. These features are progressively integrated by self-attention and cross-attention among different modalities and scales. Wang et al. [99] realized the feature enhancement and fusion using a transformer decoder. Multi-head self-attention is used to refine the initial fused feature and enhance the feature at every scale. MTFNet [100] learns intra-modality and inter-modality communication simultaneously by using a self-attention and cross-attention transformer in the encoder section.

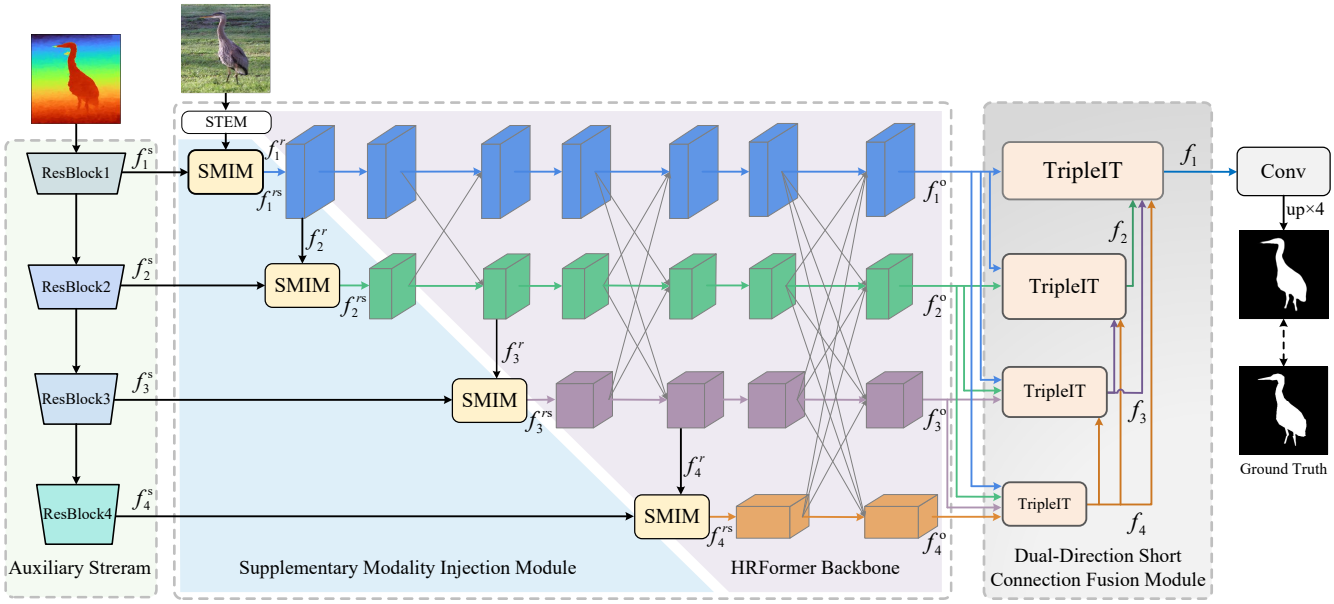


Fig. 3. Take as an example RGB-D salient object detection to show the network architecture of the proposed HRTransNet which is also applied in RGB-T and light field salient object detection. It consists of auxiliary stream, supplementary modality injection module, HRFormer backbone, and dual-direction short connection fusion module.

Transformer-based backbones have exhibited superior performance. The Visual Saliency Transformer (VST) [101] divides an image into patches and uses T2T-ViT [102] to propagate long-range dependencies between image patches. VST also designs a reverse T2T decoder simultaneously. In TransformerSOD [103], a swin transformer is used as the backbone, and a generative adversarial network and difficulty-aware learning are used to produce saliency predictions. SwinNet [27] also uses a Swin transformer as the backbone, which is used in combination with edge information to perform decoding. Zhang et al. [104] use maximum likelihood estimation to train a generative vision transformer network.

Unlike existing methods, HRFormer with high-resolution representation is used in this study to achieve two-modality SOD. The advantages of both the high-resolution network and the transformer are fully exploited in this effective combination.

III. Proposed Method

A. Motivation

The high-resolution network (e.g., HRNet and HRFormer) maintains high-resolution representation throughout forwarding propagation, which is friendly to SOD task in which the input and output have the same resolution. Therefore, we attempt to apply HRFormer to SOD. However, two-modality SOD involves the fusion of two modalities. The supplementary modality of the "depth image", "thermal image" and "focal stack" can provide the primary modality of the "color image" with spatial, thermal infrared, and focus information, but have some defects (e.g., noise, local blurring, and redundancy). Thus, we design a module that injects supplementary modalities into the primary color modality by formulating the weight for each supplementary modality from a global perspective and emphasizing the coordinate-wise spatial position of the supplementary modality. In addition, HRFormer improves upon HRNet by replacing convolution blocks with transformer blocks without changing the exchange unit and final multi-resolution fusion. However, multi-scale information is vital in two-modality SOD for capturing the details of an object. Consequently, we design another important module that has four transformer-like fusion units with dual short connections. Each transformer-like fusion unit is called an intra-feature and inter-feature interactive transformer (TripleIT). Each TripleIT takes an output resolution feature of HRFormer as the primary feature, takes higher resolution output features of HRFormer and lower resolution output features of TripleIT as associate features, and models intra- and inter- feature correlation. The designed module enhances multi-scale feature fusion and thereby the details of salient objects. In Fig. 3, RGB-D SOD is used as an example to illustrate the network architecture of HRTransNet, which can also be applied to RGB-T and LF SOD. Note that the depth map adopts Turbo colormap[1] for the better visualization. The model comprises four components, an auxiliary stream, a supplementary modality injection module, an HRFormer backbone, and a dual-direction short connection fusion module, which are described in the following sections.

B. Backbone network for primary color modality

The HRFormer [13] is an effective backbone of the primary color modality, because it can maintain high-resolution representations throughout the network that accords with the SOD task in which the input and output have the same resolution.

Therefore, we use HRFormer to extract rich and productive visual representations. HRFormer has a high-resolution convolution stem and a main body with three progressive high-to-low resolution stages. Each stage of the main body contains a

[1]https://ai.googleblog.com/2019/08/turbo-improved-rainbow-colormap-for.html

sequence of transformer blocks and exchange units. The transformer blocks perform feature updates at the same resolution, and the exchange units perform information exchange across the different resolutions. Unlike HRNet [11], HRFormer uses transformer blocks to enlarge receptive fields and achieve a global perspective. The exchange units in HRFormer are not modified.

For the purpose of clarity in the following discussion of fusion with a supplementary modality, we define the initial color features for the four stages of HRFormer as $F^r = \{f_i^r\}_{i=1}^4$, which are shown in Fig. 3.

C. Auxiliary stream for supplementary modality

Supplementary modality in RGB-D, RGB-T, and LF SOD is depth image, thermal image and focal stacks, respectively. To find discriminative and complementary information to form the supplementary to primary color modality, we use an auxiliary stream to extract supplementary features. Considering the performance and computation cost, ResNet18 [105] is selected as the auxiliary backbone. The detailed analysis can be seen in Section IV-E2. The extracted hierarchical supplementary features are denoted as $F^s = \{f_i^s\}_{i=1}^4$, where i denotes layer number.

D. Supplementary modality injection module

To use the high-resolution merit of HRFormer, supplementary modality is injected into primary color modality at the beginning of each stage. Fig. 4 shows the detail of the proposed supplementary modality injection module (SMIM). It achieves two-modality fusion between initial color features $f_i^r\ (i = 1, \cdots, 4)$ of primary color modality and supplementary features $f_i^s\ (i = 1, \cdots, 4)$ of another supplementary modality.

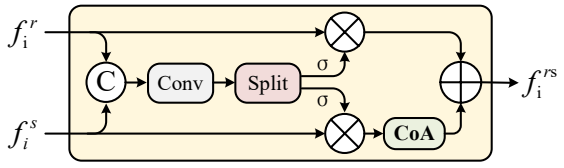


Fig. 4. Supplementary modality injection module (SMIM).

To be specific, primary color feature f_i^r and supplementary features f_i^s are first concatenated and convoluted to form a two-channel feature. Then it is split to represent two modalities. Next, sigmoid activation function and average pooling operation are successively applied to generate the weight of each modality. The process can be described as:

$$[w_i^r, w_i^s] = \mathit{Avg}\left(\mathit{sigmoid}\left(\mathit{split}\left(\mathit{Conv}\left(\mathit{Cat}\left(f_i^r, f_i^s\right)\right)\right)\right)\right) \quad (1)$$

where $\mathit{Cat}(\cdot)$ is concatenation operation, $\mathit{Conv}(\cdot)$ is a convolution operation to generate a two-channel feature, $\mathit{split}(\cdot)$ is a split operation along channel direction to yield two single-channel maps, $\mathit{sigmoid}(\cdot)$ is sigmoid activation operation, and $\mathit{Avg}(\cdot)$ is average pooling operation. Then two weights w_i^r and w_i^s are generated. They reflect the importance of each modality feature on the final result.

Next, the weights $[w_i^r, w_i^s]$ are reassigned to color and supplementary features by element-wise multiplication operation. Especially, supplementary modality further applies a coordinate attention [106] to focus on more attentive region by its direction-aware and position-sensitive ability. As a result, the noises are suppressed in supplementary modality and features are purified more discriminative. Finally, two modalities are combined by element-wise addition operation to produce new initial features $F^{rs} = \{f_i^{rs}\}_{i=1}^4$. The process can be described as:

$$f_i^{rs} = w_i^r \times f_i^r + \mathit{CoA}\left(w_i^s \times f_i^s\right) \quad (2)$$

where "+" is element-wise addition operation, "$\times$" is element-wise multiplication operation, and $\mathit{CoA}(\cdot)$ is coordinate attention [106].

E. Dual-direction short connection fusion module

HRFormer applies transformer blocks to enlarge receptive field of fused feature F^{rs}, and uses exchange units to absorb the merits of multi-scales features. The process is described as:

$$F^o = \mathit{HRFormer}\left(F^{rs}\right) \quad (3)$$

After HRFormer, output features $F^o = \{f_i^o\}_{i=1}^4$ achieve information exchange and optimization in both global and local scopes.

Next, aggregating information from output features $f_i^o (i = 1, \cdots, 4)$ of HRFormer is an essential operation for dense prediction tasks. The feature concatenation dominates the choice of aggregation operations, but its expressiveness is limited. Therefore, we devise an dual-direction short connection fusion module to aggregate all the multi-resolution output features of HRFormer to generate decoding features $F = \{f_i\}_{i=1}^4$. It consists of four Intra-feature and Inter-feature Interactive Transformers (TripleIT) $\mathbb{T} = \{T_i\}_{i=1}^4$. In the left of module, from top to bottom, high-resolution features are appended to all the low-resolution features. For example, f_1^o is added to T_2, T_3, T_4; f_2^o is added to T_3, T_4; f_3^o is added to T_4. In the right of module, from bottom to top, decoding features f_i are added to all the TripleIT $T_j (j = i-1, \cdots, 1)$ with higher resolution. For example, f_4 is added to T_3, T_2, T_1; f_3 is added to T_2, T_1; f_2 is added to T_1. Therefore, each TripleIT T_i is denoted as:

$$f_i = T_i\left(f_i^o, f_i^{asso}\right)(i = 1, \cdots, 4) \quad (4)$$

where $\boldsymbol{f_i^o}$ serves as primary feature of T_i, the rest of input features serve as associated features f_i^{asso} that are defined as:

$$f_i^{asso} = \begin{cases} \mathit{FlatCat}\left(f_2, f_3, f_4\right), & i = 1 \\ \mathit{FlatCat}\left(f_1^o, f_3, f_4\right), & i = 2 \\ \mathit{FlatCat}\left(f_1^o, f_2^o, f_4\right), & i = 3 \\ \mathit{FlatCat}\left(f_1^o, f_2^o, f_3^o\right), & i = 4 \end{cases} \quad (5)$$

where $\mathit{FlatCat}(\cdot, \cdot, \cdot)$ flattens all the features and concatenates them in the patch direction.

Each primary feature is first optimized by long-range dependency of itself, and then comprehensively combined with

associated features, to generate decoding features. The details are discussed in the following section.

At last, f_1 is convoluted and 4×upsampled to form saliency map.

F. Intra-feature and inter-feature interactive transformer

Intra-feature and inter-feature interactive transformer (TripleIT) models the correlation of intra-feature and inter-feature, achieving the optimization of primary feature based on itself and associated features. Take as an example TripleIT T_2, which is shown in Fig. 5.

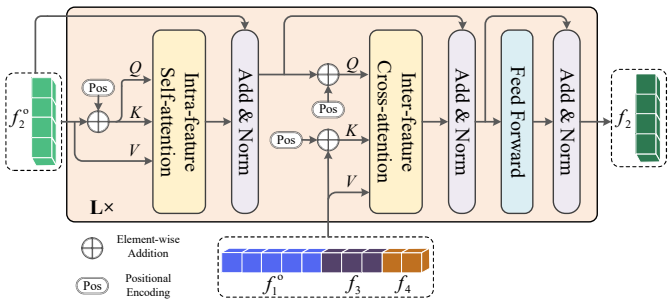


Fig. 5. Intra-feature and inter-feature interactive transformer (TripleIT).

At first, all the features with different resolutions are divided into many tokens, and each token is 1×1×128 size. Second, an intra-feature self-attention (SA) layer is applied on primary feature. It will comprehensive explore intra-feature relation with global receptive field. Third, an inter-feature cross-attention layer is used between self-boosting primary feature and its associate features. It will further excavate inter-feature relation and obtain long-range dependency across multiple resolutions. Finally, a feed-forward (FF) layer is applied. The process can be described as:

$$T_i = FF\left(CA\left(SA\left(f_i^o\right), f_i^{asso}\right)\right) \tag{6}$$

where $SA(\cdot)$ is intra-feature self-attention, $CA(\cdot)$ is inter-feature cross-attention, f_i^{asso} is associated features of f_i^o and $FF(\cdot)$ is feedforward network (FFN). Each operation follows by a residual connection and normalization layer.

The intra-feature self-attention uses the primary feature as input Q, K, and V, and output self-boosting primary feature. The inter-feature cross-attention adopts self-boosting primary feature as Q, meanwhile uses the associated features as K and V. The intra-feature self-attention and inter-feature cross-attention are both implemented by efficient attention [107] for less memory and computational costs. In the inter-feature cross-attention layer, since the length of patches of primary feature and associated features are different, position embedding [108] is used to avoid the position errors between tokens for learning the attention weight.

By stacking $L = 2$ above architectures, the primary feature is refined based on its intra-feature self-relationship, and further optimized by useful information of all the associated features.

IV. Experiments

A. Datasets

RGB-D SOD dataset includes NLPR [109] with 1,000 images, NJU2K [110] with 1,985 images, STERE [111] with 1,000 images, DES [112] with 135 images, SIP [59] with 929 images, and DUT [113] with 1,200 images. **RGB-T SOD dataset** includes VT821 [72], VT1000 [74], and VT5000 [114]. **LF SOD dataset** includes LFSD [81] with 100 light field data, HFUT-Lytro [115] with 255 samples, DUTLF-FS [84] with 1,462 light field images.

B. Evaluation metrics

Precision-recall (PR) curve [116] reflects the relation of precision and recall. The best model has both the best accuracy and the best recall. **S-measure [117]** evaluates structural similarity and emphasizes structural integrity. **F-measure [118]** is the adaptive value presentation of the PR curve. **E-measure [119]** captures adaptive global and local similarity. **Mean absolute error (MAE) [120]** measures per-pixel absolute difference.

C. Implementation details

We use an NVIDIA RTX 3090 GPU to train our model. The training set of RGB-D SOD comprises 2,185 paired RGB and depth images from NJU2K and NLPR. When testing the DUT dataset, our training set adds the DUT training set with 800 image pairs. The training set of RGB-T SOD includes 2,500 image pairs in VT5000. The training set of LF SOD consists of 100 samples in HFUT-Lytro and 1,000 samples in DUTLF-FS. The rest samples involved in Section IV-A are tested. HRFormer in the primary stream and ResNet18 in the auxiliary stream use pre-trained parameters. The rest of parameters are initialized to PyTorch default settings. The loss function adopts pixel position aware loss [121]. During training, data enhancement, i.e., random flipping, rotating and border clipping, are conducted. The size of input image is set as 224×224. The batch size is 15 and the initial learning rate is 5e-5. The model converges within 300 epochs.

D. Comparisons with SOTAs

1) RGB-D SOD: There are many RGB-D SOD methods which surpass previous methods in the last year. Therefore, we compare our model with these pioneering works published in 2021, including JL-DCF [122], CDNet [123], DPANet [69], HAINet [124], DSNet [46], CCAFNet [125], EBFSP [126], MMNet [77], RD3D [61], TriTransNet [97], DCF [56], DSA2F [63], VST [101], SPNet [127], EBMG [104], and SwinNet [27].

Quantitative Evaluation. The PR curve in Fig. 6 shows accuracy and recall of different models. Our curves are better than the others on all datasets. It benefits from the initial selective fusion of two modalities and the integration of multi-resolution output features of HRFormer. Four evaluation metrics in Table. I also give the same results when compared with pioneering works published in CVPR, ICCV, NeurIPS, and so on. Most evaluation metrics are superior to the others.

It indicates that our method is better than the others, and comparable with EBMG [104] and SwinNet [27]. Moreover, from the bottom of the table, we can find our method is superior in the computation cost. It has the fewer parameters and lower FLOPs than most compared methods.

Qualitative Evaluation. The visual examples in Fig. 7 show the performance of our model in some scenes: indistinguishable foreground object (1^{st}-2^{nd} rows), complex scene (3^{rd}-4^{th} rows), low-quality depth image (5^{th}-6^{th} rows), small objects (7^{th}-8^{th} rows), multiple objects (9^{th}-10^{th} rows), and fine-grained objects (11^{th}-12^{th} rows). It is due to the proposed several modules. For example, supplementary modality injection module reduces the influence of low-quality depth images by effectively filtering out depth noise. It ensures good performance in the scene with low-quality depth image. For example, the dual-direction short connection fusion module combines the merits of multi-resolution features, showing superior performance in detecting fine-grained details. For example, HRFormer with transformer blocks maintains larger receptive fields than the original HRNet. It ensures powerful detection ability in small objects and multiple objects, complex scenes, etc.

2) RGB-T SOD: Compared with RGB-D SOD, RGB-T SOD is developed slower. Comparison methods include MTMR [72], M3S-NIR [71], SGDL [74], ADF [114],MIDD [76], ECFFNet [75], MMNet [77], CSRNet [79], CGFNet [78], and SwinNet [27], which are published in recent five years.

Quantitative Evaluation. The PR curve in Fig. 8 shows that our method wins the others with a great margin. Four evaluation metrics in Table. II also reveal the outstanding performance of our models. The figure and table both indicate our method is robust when applied in RGB and thermal image pairs.

Qualitative Evaluation. The visual examples in Fig. 9 show the performance comparison of different models in some scenes: indistinguishable foreground object (1^{st} row), cluttered scene (2^{nd} row), weak light condition (3^{rd} row), low-quality thermal image (4^{th} row), small objects (5^{th} row), multiple objects (6^{th} row), and scenes filled with noises (7^{th} row). They suggest the effectiveness of proposed modules in the SOD task by the combination of color and thermal infrared information.

3) LF SOD: Compared with RGB-D and RGB-T SOD, LF SOD is studied by fewer researchers. Comparison methods comprises seven methods published in recent four years, including MoLF [86], DLFS [128], LFNet [85], ERNet [83], SA-Net [87], DLGLRG [88], PANet [129]. As we all know, LF image includes multi-focal, multi-view [130], and EPIs [10], [131]. We mainly discuss multi-focal LF whose input is an all-in-focus image and focal stack with not more than 12 focal slices. When the number of focal slices is less than 12, we use an all-zero image to pad [87]. Then all-in-focus image serves as the primary color feature, and multiple slices are concatenated to serve as the supplementary feature.

Quantitative Evaluation. The PR curve in Fig. 10 shows that our model outperforms the others in HFUT-Lytro and DUTLF-FS datasets, and is slightly better than the others in LFSD dataset. Four evaluation metrics in Table. III shows a performance improvement. The excellent performance is achieved only by concatenating focal slices. It also verifies the strong compatibility of our model in processing the two-modality SOD tasks.

Qualitative Evaluation. The visual examples in Fig. 11 show the performance of our proposed model in some challenging cases: big objects (1^{st} row), small objects (2^{nd} row), multiple objects (3^{rd} row), similar foreground and background (4^{th} row), and complex scenes (5^{th} row). It suggests that our method can model long-range dependency and detect salient objects from a global perspective, meanwhile maintaining fine-grained detail of the local region.

E. Ablation studies

We take as an example RGB-D SOD to conduct the ablation study about the selection of primary stream backbone and auxiliary stream backbone, effectiveness of supplementary modality injection module and dual-direction short connection fusion module.

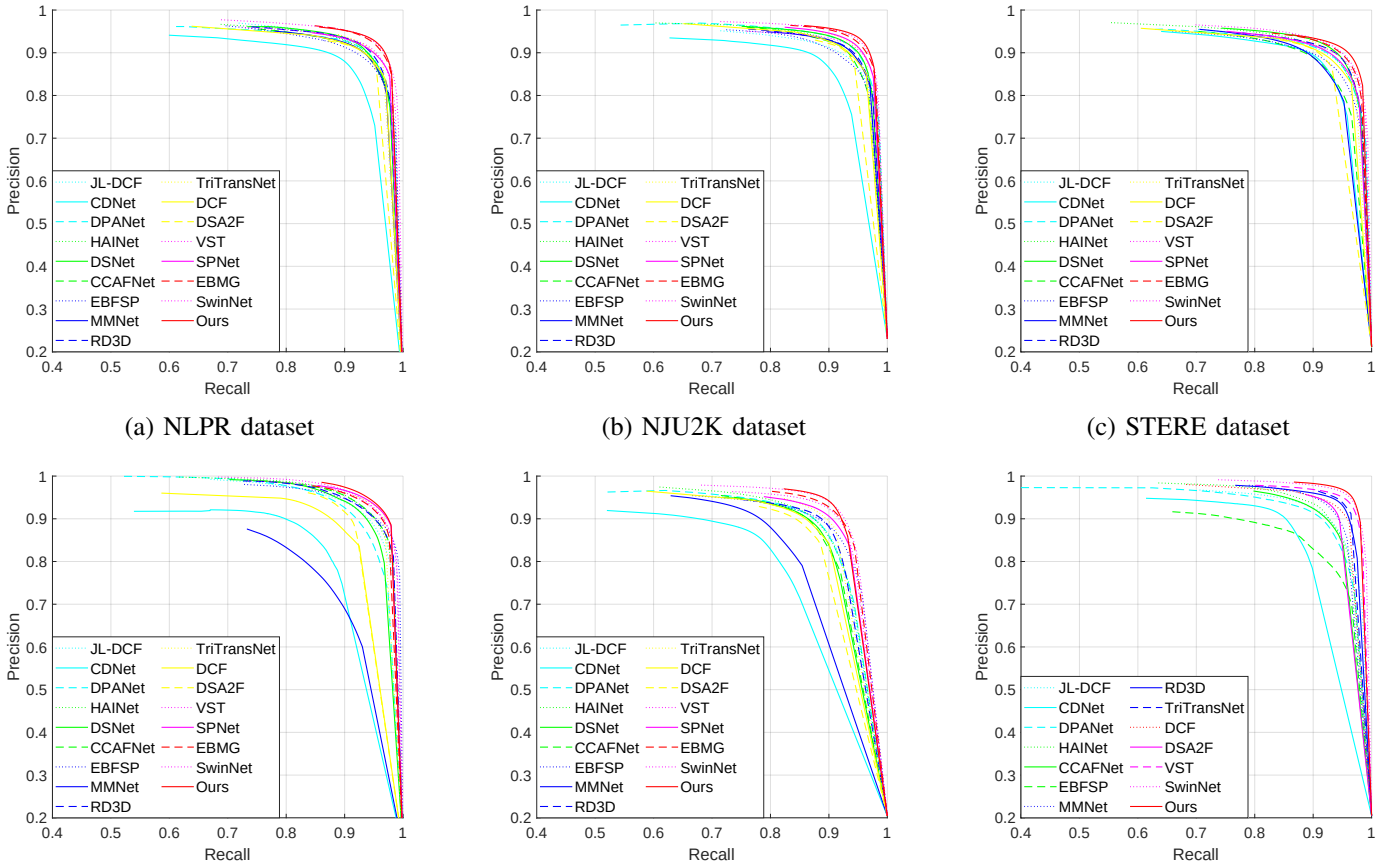
1) Selection of primary stream backbone: Some backbones are selected as feature extractors for primary stream. They are ResNet [105], Swin Transformer [132], HRNet [11], HRFormer [13]. Table. IV shows the results which indicate that HRFormer plays an important role in improving the performance. Moreover, the model with HRFormer backbone has the less computation cost.

2) Selection of auxiliary stream backbone: Some backbones are selected as feature extractors for auxiliary stream. They are ResNet18 [105], ResNet50 [105], HRFormer [13], Swin Transformer [132], Segformer [133], ConvNet [134]. Table. V shows the results which indicate that ResNet18 can achieve comparable performance with a fewer parameters and FLOPs. Therefore, the final model adopts ResNet18 as the auxiliary stream backbone.

3) Effectiveness of modules: To offer deeper insights into supplementary modality injection module (SMIM) and the dual-direction short connection fusion module with four TripleITs, we perform the ablation study. Fig. 12 shows the baseline model and our model. The baseline model replaces SMIM with addition, and replaces four TripleITs with progressive decoding process with the addition and upsampling.

From Table. VI, we can see that all the evaluation values are improved except for S-measure and MAE in STERE dataset when adding SMIM to the baseline model. It illustrates that SMIM plays a positive role in improving the performance. It benefits from the weighted fusion of two-modality data and the special coordinate attention assigned on auxiliary features with the latent noise. Furthermore, TripleITs achieve the impressive improvement in all the evaluation metrics. It illustrates the obvious effect of four TripleITs by absorbing the advantage of all the features with different resolutions to achieve complementary fusion and comprehensive detail enhancement. Finally, the model equipped with SMIM and TripleITs achieves the best performance.

We also present some visual comparisons in Fig. 13. From the comparison of (a) and (c), we discover that addition



(a) NLPR dataset (b) NJU2K dataset (c) STERE dataset

(d) DES dataset (e) SIP dataset (f) DUT dataset

Fig. 6. The comparison of P-R curves on six RGB-D datasets.

TABLE I

THE COMPARISON OF S-MEASURE, ADAPTIVE F-MEASURE, ADAPTIVE E-MEASURE, MAE OF DIFFERENT RGB-D SOD MODELS. "-" MEANS THAT THE METHOD DOES NOT RELEASE TEST RESULTS FOR THIS DATASET OR CODE.

Datasets	Metric	JL-DCF TPAMI21 [122]	CDNet TIP21 [123]	DPANet TIP21 [69]	HAINet TIP21 [124]	DSNet TIP21 [46]	CCAFNet TMM21 [125]	EBFSP TMM21 [126]	MMNet TCSVT21 [77]	RD3D AAAI21 [61]	TriTransNet ACM MM21 [97]	DCF CVPR21 [56]	DSA2F CVPR21 [63]	VST ICCV21 [101]	SPNet ICCV21 [127]	EBMG NeurIPS21 [104]	SwinNet TCSVT21 [27]	HRTransNet Ours
NLPR	$S\uparrow$	.925	.902	.928	.924	.926	.922	.915	.925	.933	.928	.923	.918	.931	.927	.938	.941	**.942**
	$F_\beta\uparrow$	.878	.848	.875	.897	.886	.880	.897	.889	.886	.909	.876	.892	.886	.904	**.920**	.908	.919
	$E_\xi\uparrow$	.954	.935	.953	.957	.955	.951	.952	.950	.958	.960	.950	.950	.954	.958	.967	.967	**.969**
	MAE$\downarrow$	.022	.032	.024	.024	.024	.026	.026	.024	.022	.020	.024	.024	.023	.021	.018	.018	**.016**
NJU2K	$S\uparrow$	.903	.885	.921	.912	.921	.910	.903	.911	.928	.920	.912	.904	.922	.924	.929	.935	**.933**
	$F_\beta\uparrow$	.884	.866	.887	.900	.907	.897	.894	.900	.910	.919	.902	.898	.899	.917	.923	.922	**.928**
	$E_\xi\uparrow$	.913	.908	.915	.922	.920	.920	.907	.919	.920	.925	.924	.922	.914	.932	**.935**	.934	.931
	MAE$\downarrow$	.041	.048	.035	.038	.034	.037	.039	.038	.033	.030	.035	.039	.034	.028	.028	.027	**.026**
STERE	$S\uparrow$	.903	.896	.909	.907	.915	.892	.900	.891	.914	.908	.902	.897	.913	.907	.916	.919	**.921**
	$F_\beta\uparrow$	.869	.873	.875	.885	.894	.869	.870	.880	.881	.893	.884	.893	.878	.888	.899	.893	**.904**
	$E_\xi\uparrow$	.918	.922	.920	.925	.929	.921	.912	.924	.921	.927	.929	.927	.917	**.930**	**.930**	.929	**.930**
	MAE$\downarrow$	.040	.042	.040	.040	.036	.044	.045	.045	.039	.033	.039	.039	.038	.037	.032	.033	**.030**
DES	$S\uparrow$	.931	.876	.919	.935	.927	.938	.937	.830	.950	.943	.904	.916	.943	.944	.938	.945	**.947**
	$F_\beta\uparrow$	.900	.840	.900	.924	.910	.915	.913	.746	.915	.936	.876	.901	.917	.935	.928	.926	**.938**
	$E_\xi\uparrow$	.969	.921	.963	.974	.970	.975	.974	.893	.979	.981	.950	.955	.979	.983	.977	.980	**.983**
	MAE$\downarrow$	.020	.034	.023	.018	.021	.018	.018	.058	.017	**.014**	.024	.023	.017	**.014**	.016	.016	**.014**
SIP	$S\uparrow$	.880	.823	.883	.880	.876	.877	.885	.836	.892	.886	.875	.862	.904	.894	.902	**.911**	.909
	$F_\beta\uparrow$	.873	.805	.865	.875	.865	.864	.869	.839	.883	.892	.875	.865	.895	.893	.908	.912	**.916**
	$E_\xi\uparrow$	.921	.880	.921	.919	.910	.915	.917	.882	.924	.924	.920	.908	.937	.930	.938	**.943**	**.943**
	MAE$\downarrow$	.049	.076	.051	.053	.052	.054	.049	.075	.046	.043	.052	.057	.040	.043	.037	**.035**	**.035**
DUT	$S\uparrow$	.906	.880	.899	.909	-	.904	.858	.920	.936	.934	.924	.921	.943	-	-	.949	**.951**
	$F_\beta\uparrow$	.882	.874	.881	.905	-	.903	.841	.919	.925	.936	.925	.926	.931	-	-	.944	**.955**
	$E_\xi\uparrow$	.931	.918	.923	.937	-	.940	.890	.951	.953	.957	.952	.950	.960	-	-	.968	**.972**
	MAE$\downarrow$	.043	.048	.048	.038	-	.037	.067	.032	.030	.025	.030	.030	.025	-	-	.020	**.018**
Params(M)$\downarrow$		143.5	**32.9**	92.4	59.8	-	-	-	64.1	46.9	138.7	108.5	36.5	83.0	67.9	88.8	198.7	58.9
FLOPs(G)$\downarrow$		861.2	72.1	58.9	181.4	-	-	-	42.4	26.8	292.3	54.0	364.4	31.0	150.3	303.9	124.3	**17.1**

TABLE II
THE COMPARISON OF S-MEASURE, ADAPTIVE F-MEASURE, ADAPTIVE E-MEASURE, MAE OF DIFFERENT RGB-T SOD MODELS.

Datasets	Metric	MTMR [72] IGTA18	M3S-NIR [71] MIPR19	SGDL [74] TMM19	ADF [114] TMM22	MIDD [76] TIP21	ECFFNet [75] TCSVT21	MMNet [77] TCSVT21	CSRNet [79] TCSVT21	CGFNet [78] TCSVT21	SwinNet [27] TCSVT21	HRTransNet Ours
VT821	$S\uparrow$	.725	.723	.765	.810	.871	.877	.875	.885	.881	.904	**.906**
	$F_\beta\uparrow$	.662	.734	.730	.716	.804	.810	.798	.830	.845	.847	**.853**
	$E_\xi\uparrow$	.815	.859	.847	.842	.895	.902	.893	.908	.912	.926	**.929**
	MAE↓	.108	.140	.085	.077	.045	.034	.040	.038	.038	.030	**.026**
VT1000	$S\uparrow$	.706	.726	.787	.910	.915	.923	.917	.918	.923	**.938**	**.938**
	$F_\beta\uparrow$	.715	.717	.764	.847	.882	.876	.863	.877	**.906**	.896	.900
	$E_\xi\uparrow$	.836	.827	.856	.921	.933	.930	.924	.925	.944	**.947**	.945
	MAE↓	.119	.145	.090	.034	.027	.021	.027	.024	.023	.018	**.017**
VT5000	$S\uparrow$	.680	.652	.750	.863	.867	.874	.864	.868	.883	**.912**	**.912**
	$F_\beta\uparrow$	.595	.575	.672	.778	.801	.806	.785	.810	.851	.865	**.871**
	$E_\xi\uparrow$	.795	.780	.824	.891	.897	.906	.890	.905	.922	.942	**.945**
	MAE↓	.114	.168	.089	.048	.043	.038	.043	.042	.035	.026	**.025**

TABLE III
THE COMPARISON OF S-MEASURE, ADAPTIVE F-MEASURE, ADAPTIVE E-MEASURE, MAE OF DIFFERENT LF SOD MODELS.

Datasets	Metric	MoLF [86] NeurIPS19	DLFS [128] IJCAI19	LFNet [85] TIP20	ERNet [83] AAAI20	SA-Net [87] BMVC21	DLGLRG [88] ICCV21	PANet [129] TCYB21	HRTransNet Ours
LFSD	$S\uparrow$	.830	.735	.806	.835	.841	.867	.849	**.875**
	$F_\beta\uparrow$	.819	.713	.793	.839	.845	.861	.849	**.875**
	$E_\xi\uparrow$	.886	.805	.870	.887	.889	**.898**	.893	.896
	MAE↓	.089	.149	.101	.082	.074	.069	.076	**.056**
HFUT-Lytro	$S\uparrow$	.742	.741	.781	.778	.784	.766	.795	**.827**
	$F_\beta\uparrow$	.627	.616	.659	.705	.736	.709	.724	**.774**
	$E_\xi\uparrow$	.785	.784	.808	.831	.850	.841	.851	**.869**
	MAE↓	.095	.097	.076	.082	.078	.071	.074	**.062**
DUTLF-FS	$S\uparrow$	.887	.841	.882	.900	.918	.928	.908	**.938**
	$F_\beta\uparrow$	.843	.801	.842	.888	.920	.923	.897	**.944**
	$E_\xi\uparrow$	.923	.891	.914	.942	.954	.952	.940	**.962**
	MAE↓	.052	.076	.054	.040	.032	.031	.039	**.024**

TABLE IV
ABLATION STUDY ABOUT SELECTION OF PRIMARY MODALITY BACKBONE.

Variant	Params(M)↓	FLOPs(G)↓	NLPR				NJU2K				STERE				SIP			
			$S\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	MAE↓	$S\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	MAE↓	$S\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	MAE↓	$S\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	MAE↓
ResNet [105]	67.7	22.3	.909	.865	.947	.029	.882	.864	.903	.051	.892	.862	.913	.045	.849	.835	.908	.064
Swin Transformer [132]	119.9	23.7	.908	.866	.949	.028	.899	.879	.916	.041	.879	.841	.907	.049	.880	.872	.925	.048
HRNet [11]	81.2	21.5	.932	.907	.961	.020	.922	.916	.928	.031	.909	.889	.926	.036	.902	.902	.935	.038
HRFormer [13]	**58.9**	**17.1**	**.942**	**.919**	**.969**	**.016**	**.933**	**.928**	**.931**	**.026**	**.921**	**.904**	**.930**	**.030**	**.909**	**.916**	**.943**	**.035**

TABLE V
ABLATION STUDY ABOUT SELECTION OF SUPPLEMENTARY MODALITY BACKBONE.

Variant	Params(M)↓	FLOPs(G)↓	NLPR				NJU2K				STERE				SIP			
			$S\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	MAE↓	$S\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	MAE↓	$S\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	MAE↓	$S\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	MAE↓
ResNet18 [105]	**58.26**	**17.12**	**.942**	**.919**	**.969**	**.016**	**.933**	**.928**	.931	.026	.921	.904	.930	.030	.909	.916	**.943**	.035
ResNet50 [105]	82.36	22.07	**.942**	.915	**.969**	**.016**	.930	.923	.930	.027	.921	.902	**.932**	.030	**.910**	**.918**	**.943**	**.034**
HRFormer [13]	88.75	27.23	.938	.913	.965	.017	.931	.924	.930	.027	.921	.903	**.932**	.031	.907	.915	.940	.035
Swin Transformer-B [132]	137.63	31.38	.940	.916	.967	.017	.932	.927	**.933**	.026	**.923**	**.906**	931	**.029**	.909	.912	.941	**.034**
Segformer-B4 [133]	108.04	24.02	.940	.917	.966	.017	**.933**	**.928**	**.933**	**.025**	.921	.904	**.932**	.030	**.910**	**.918**	**.943**	**.034**
ConvNeXt-B [134]	138.52	31.62	.938	.914	.964	.017	.928	.921	.931	.029	.920	.902	.930	.030	.909	.915	.941	.035

TABLE VI
THE ABLATION STUDY ABOUT SMIM AND FOUR TRIPLEITS.

Variant	Candidate			NLPR				NJU2K				STERE				SIP			
	Baseline	SMIM	TripleITs	$S\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	MAE↓	$S\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	MAE↓	$S\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	MAE↓	$S\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	MAE↓
No.1	✓			.928	.892	.956	.021	.925	.911	.921	.030	.917	.893	.924	.033	.899	.888	.930	.040
No.2	✓	✓		.935	.910	.963	.018	.928	.919	.928	.028	.914	.897	.928	.034	.908	.912	.940	.036
No.3	✓		✓	.936	.909	.963	.019	.932	.926	.930	**.026**	.920	.903	**.932**	.031	**.909**	.911	.940	**.035**
No.4	✓	✓	✓	**.942**	**.919**	**.969**	**.016**	**.933**	**.928**	**.931**	**.026**	**.921**	**.904**	.930	**.030**	**.909**	**.916**	**.943**	**.035**

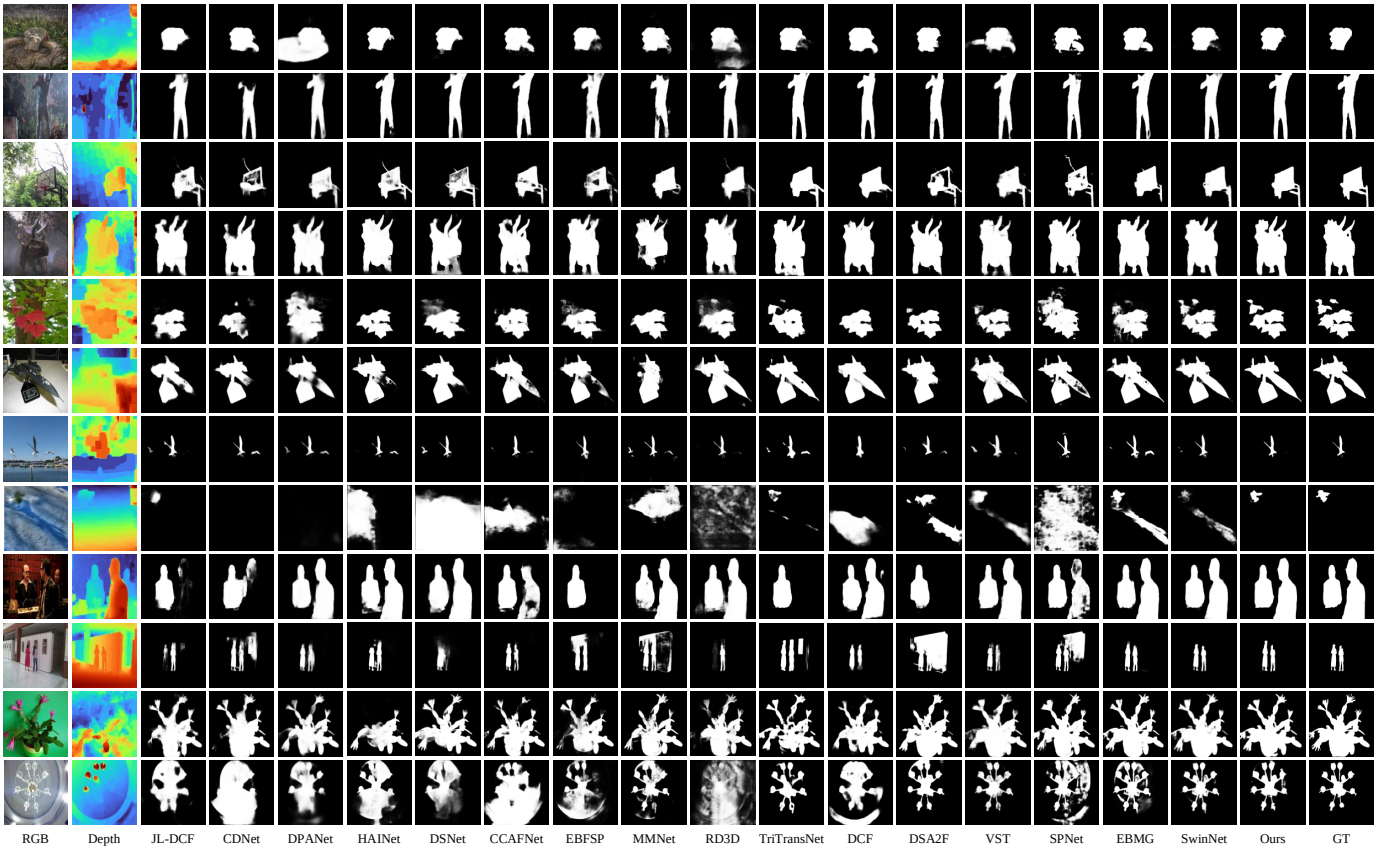
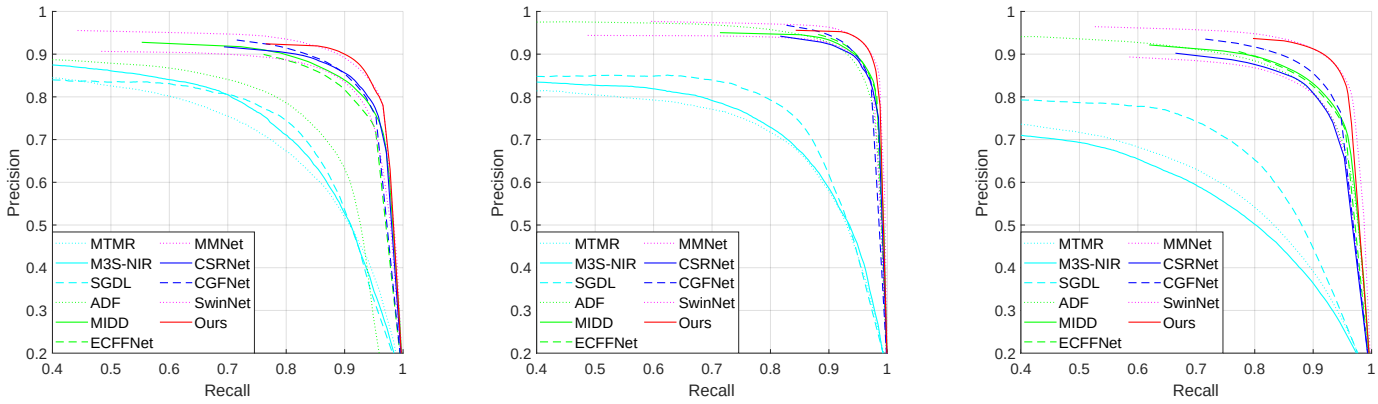


Fig. 7. Visual comparison of RGB-D SOD. Our HRTransNet is outstanding in some scenes: indistinguishable foreground object (1^{st}-2^{nd} rows), complex scene (3^{rd}-4^{th} rows), low-quality depth image (5^{th}-6^{th} rows), small objects (7^{th}-8^{th} rows), multiple objects (9^{th}-10^{th} rows), and fine-grained objects (11^{th}-12^{th} rows).



(a) VT821 dataset (b) VT1000 dataset (c) VT5000 dataset

Fig. 8. The comparison of P-R curves on three RGB-T datasets.

operation is inferior to our proposed method in the depth modality injection process, especially when depth modality is incomplete or inadequate, which is demonstrated by the first and second lines. Compared with addition, ours can assign different weight to each modality and better suppress the noise in the depth modality. Accordingly the better results can be achieved. From the comparison of (b) and (c), we find the simple decoder can locate the salient object accurately but generate blurry boundaries. It indicates the simple decoder model has no advantage in local detail representation. Compared with it, our model is better in polishing details by the long-range dependency of the transformer and achieving excellent performance with less noise.

F. Failure cases

The proposed model has a good detection performance in most cases. However, there are a few failure cases, as shown in Fig. 14. The first line can't highlight the person

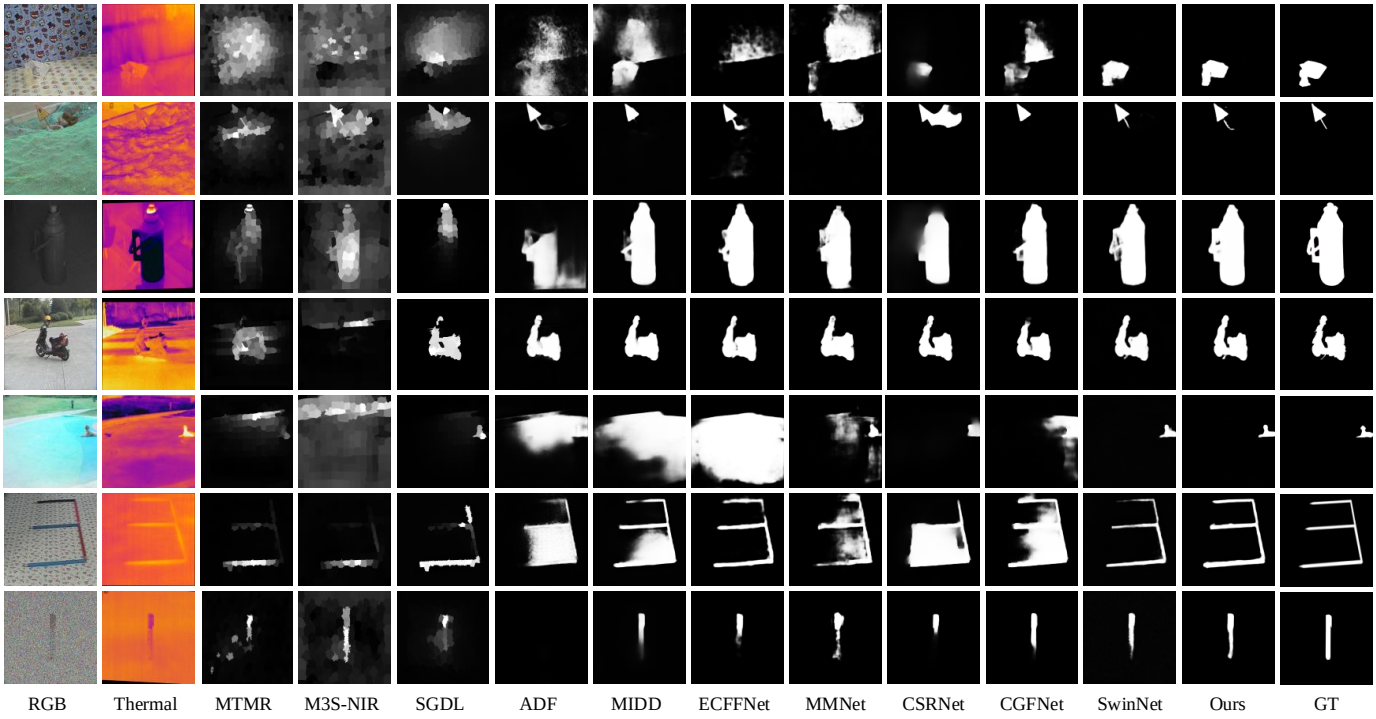
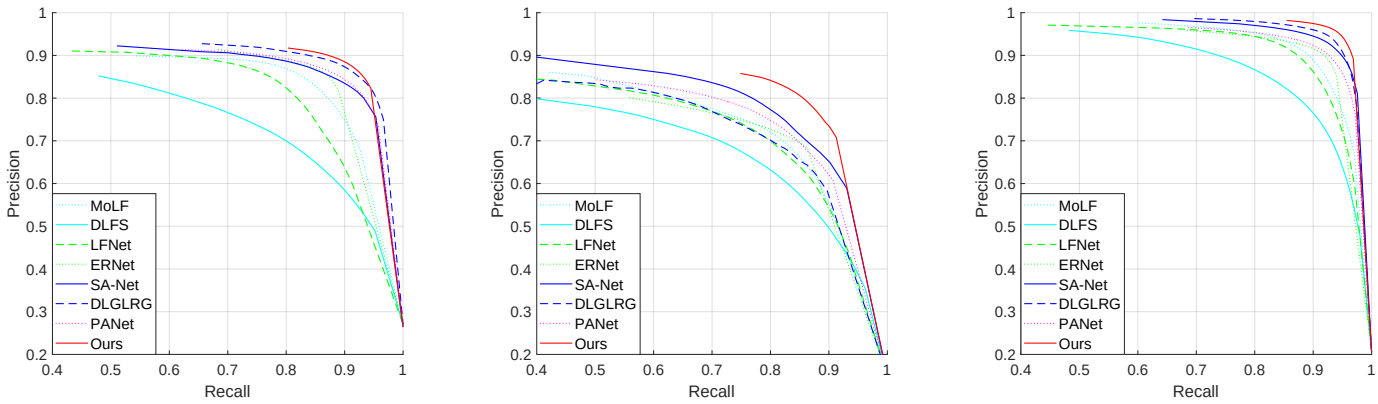


Fig. 9. Visual comparison of RGB-T SOD. Our HRTransNet is outstanding in some scenes: indistinguishable foreground object (1^{st} row), cluttered scene (2^{nd} row), weak light condition (3^{rd} row), low-quality thermal image (4^{th} row), small objects (5^{th} row), multiple objects (6^{th} row), and scenes filled with noises (7^{th} row).



(a) LFSD (b) HFUT-Lytro (c) DUTLF-FS

Fig. 10. The comparison of P-R curves on three light field datasets.

due to serious occlusions. The second line generates the error result due to extremely similar background. The propeller in third line is incomplete because it is misidentified as the background instead of the part of the glider. The similar failed detection results also occur in the other state-of-the-art RGB-D SOD methods. Therefore, salient object detection need to be further studied to solve the occlusions, extreme background interference, and object completeness problems.

V. Conclusions

In this paper, we study two-modality salient object detection tasks based on HRFormer. For the one-modality to two-modality extension, we inject a supplementary modality to effectively combine two-modality cues. We improve HRFormer by modifying the final multi-resolution fusion using intra-feature and inter-feature interactive transformers. Each transformer-like unit enhances the feature by the feature itself and the features with other resolutions. The transmission of features to TripleIT constitutes a dual-direction short connection flow. The proposed HRTransNet is applied to RGB-D, RGB-T, and LF SOD, and exhibits excellent performance. We will attempt other two-modality SOD tasks and exploit multi-modality SOD and co-saliency tasks in future studies. Furthermore, we will discuss SOD task in virtual reality and augmented reality (VR/AR) application.

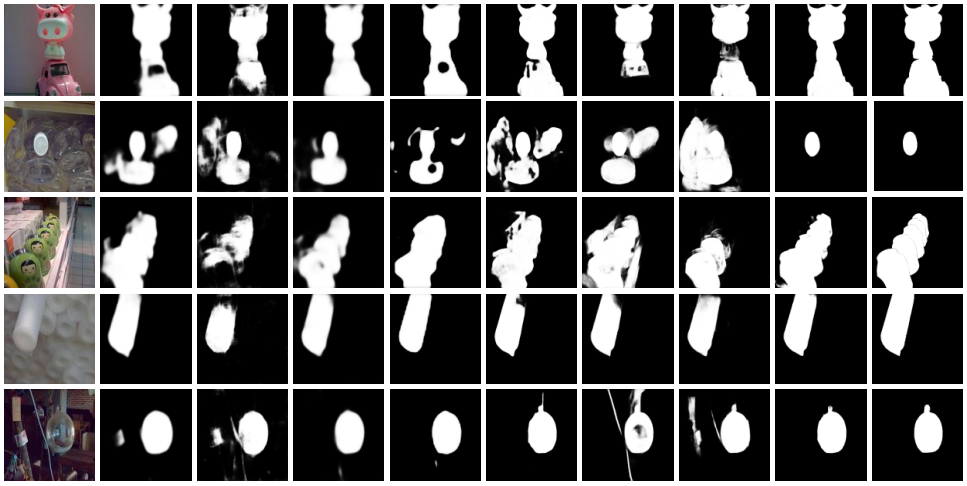


Fig. 11. Visual comparison of LF SOD. Our HRTransNet is outstanding in some challenging cases: big objects (1^{st} row), small objects (2^{nd} row), multiple objects (3^{rd} row), similar foreground and background (4^{th} row), and complex scenes (5^{th} row).

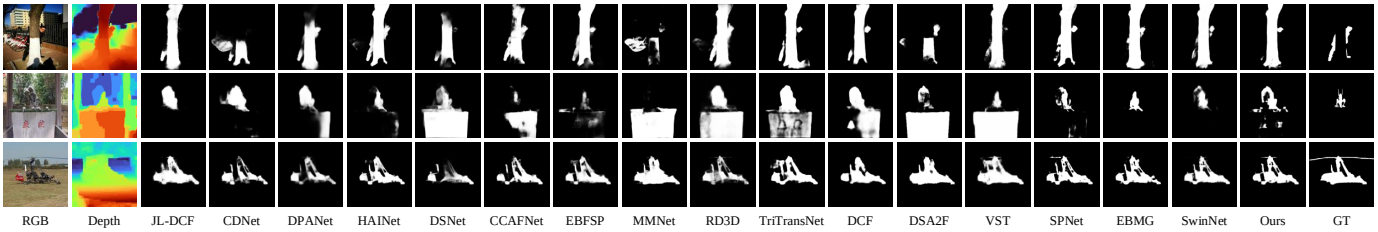
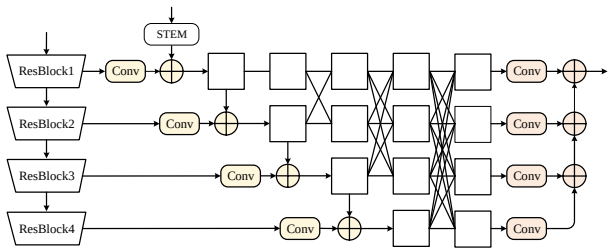
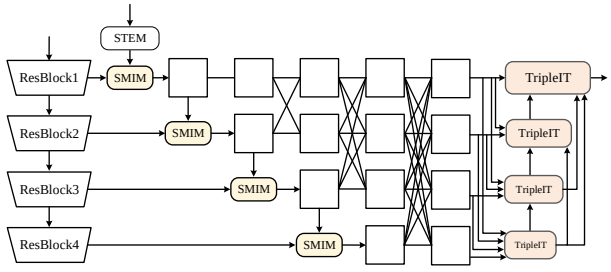


Fig. 14. Failure cases. RGB-D SOD methods all fail in occlusion (1^{st} line), extreme background interference (2^{nd} line), and object completeness (3^{rd} line).



(a) Baseline



(b) Ours

Fig. 12. The baseline and ours in the ablation study about SMIM and four TripleITs.

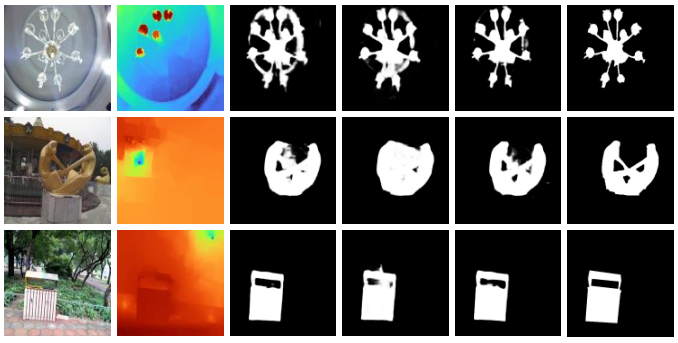


Fig. 13. Visual comparison. From the comparison between (a) and (c), we can verify the effectiveness of supplementary modality injection module. From the comparison between (b) and (c), we can prove the advantage of dual-direction short connection fusion module.

REFERENCES

[1] S. K. Yarlagadda, D. M. Montserrat, D. Güera, C. J. Boushey, D. A. Kerr, and F. Zhu, "Saliency-Aware Class-Agnostic Food Image Segmentation," *ACM Transactions on Computing for Healthcare*, vol. 2, no. 3, pp. 1–17, 2021.

[2] H. Huang, M. Cai, L. Lin, J. Zheng, X. Mao, X. Qian, Z. Peng, J. Zhou, Y. Iwamoto, X.-H. Han *et al.*, "Graph-Based Pyramid Global Context Reasoning with a Saliency-Aware Projection for COVID-19 Lung Infections Segmentation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE. IEEE, 2021, pp. 1050–1054.

[3] C. Ma, Z. Miao, X.-P. Zhang, and M. Li, "A Saliency Prior Context Model for Real-Time Object Tracking," *IEEE Transactions on Multimedia*, pp. 2415–2424, 2017.

[4] 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 Recognition*, vol. 100, p. 107130, 2020.

[5] W. Wang, J. Shen, and H. Ling, "A Deep Network Solution for Attention and Aesthetics Aware Photo Cropping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1531–1544, 2018.

[6] Y. Xu, W. Xu, M. Wang, L. Li, G. Sang, P. Wei, and L. Zhu, "Saliency Aware Image Cropping with Latent Region Pair," *Expert Systems with Applications*, vol. 171, p. 114596, 2021.

[7] M. Ahmadi, N. Karimi, and S. Samavi, "Context-Aware Saliency Detection for Image Retargeting Using Convolutional Neural Networks," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11 917–11 941, 2021.

[8] Z. Weng, W. Li, and Z. Jin, "Human Activity Prediction Using Saliency-Aware Motion Enhancement and Weighted LSTM Network," *EURASIP Journal on Image and Video Processing*, vol. 2021, no. 1, pp. 1–23, 2021.

[9] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "RGB-D Salient Object Detection: A Survey," *Computational Visual Media*, pp. 37–69, 2021.

[10] K. Fu, Y. Jiang, G.-P. Ji, T. Zhou, Q. Zhao, and D.-P. Fan, "Light Field Salient Object Detection: A Review and Benchmark," *Computational Visual Media*, pp. 1–26, 2022.

[11] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[12] J. Gu, H. Kwon, D. Wang, W. Ye, M. Li, Y.-H. Chen, L. Lai, V. Chandra, and D. Z. Pan, "Multi-Scale High-Resolution Vision Transformer for Semantic Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 094–12 103.

[13] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "HRFormer: High-Resolution Transformer for Dense Prediction," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1–15, 2021.

[14] B. Ren, S. Ma, B. Hou, and D. Hong, "Dual-Stream High Resolution Network for Multi-Source Remote Sensing Image Segmentation," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, IEEE. IEEE, 2021, pp. 3440–3443.

[15] U. Sajid, X. Chen, H. Sajid, T. Kim, and G. Wang, "Audio-Visual Transformer Based Crowd Counting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2249–2259.

[16] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-Resolution Representations for Labeling Pixels and Regions," *ArXiv Preprint ArXiv:1904.04514*, 2019.

[17] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep High-Resolution Representation Learning for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2021.

[18] C. Zhang, N. He, Q. Sun, X. Yin, and K. Lu, "Human Pose Estimation Based on Attention Multi-Resolution Network," in *Proceedings of the 2021 International Conference on Multimedia Retrieval*, 2021, pp. 682–687.

[19] H. Wu, C. Liang, M. Liu, and Z. Wen, "Optimized HRNet for Image Semantic Segmentation," *Expert Systems with Applications*, vol. 174, p. 114532, 2021.

[20] F. Akhyar, C.-Y. Lin, and G. S. Kathiresan, "A Beneficial Dual Transformation Approach for Deep Learning Networks Used in Steel Surface Defect Detection," in *Proceedings of the 2021 International Conference on Multimedia Retrieval*, 2021, pp. 619–622.

[21] J. Shao, B. Du, C. Wu, M. Gong, and T. Liu, "HRSiam: High-Resolution Siamese Network, Towards Space-Borne Satellite Video Tracking," *IEEE Transactions on Image Processing*, vol. 30, pp. 3056–3068, 2021.

[22] W. Wang, J. Zhang, L. Niu, H. Ling, X. Yang, and L. Zhang, "Parallel Multi-Resolution Fusion Network for Image Inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 559–14 568.

[23] X. Wu, T.-Z. Huang, L.-J. Deng, and T.-J. Zhang, "Dynamic Cross Feature Fusion for Remote Sensing Pansharpening," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 687–14 696.

[24] X. Cai, B. Chen, J. Zeng, J. Zhang, Y. Sun, X. Wang, Z. Ji, X. Liu, X. Chen, and S. Shan, "Gaze Estimation with an Ensemble of Four Architectures," *ArXiv Preprint ArXiv:2107.01980*, 2021.

[25] C. Yu, B. Xiao, C. Gao, L. Yuan, L. Zhang, N. Sang, and J. Wang, "Lite-Hrnet: A Lightweight High-Resolution Network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 440–10 450.

[26] M. Ding, X. Lian, L. Yang, P. Wang, X. Jin, Z. Lu, and P. Luo, "HR-NAS: Searching Efficient High-Resolution Neural Architectures with Lightweight Transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2982–2992.

[27] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "SwinNet: Swin Transformer Drives Edge-Aware RGB-D and RGB-T Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–13, 2021.

[28] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "CSwin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 124–12 134.

[29] H. Yang, L. Guo, X. Wu, and Y. Zhang, "Scale-Aware Attention-Based Multi-Resolution Representation for Multi-Person Pose Estimation," *Multimedia Systems*, pp. 1–11, 2021.

[30] X. Zhao, C. Guo, and Q. Zou, "Human Pose Estimation with Gated Multi-Scale Feature Fusion and Spatial Mutual Information," *The Visual Computer*, pp. 1–19, 2021.

[31] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[32] M. Ding, S. Zhang, and J. Yang, "Learning a Dynamic High-Resolution Network for Multi-Scale Pedestrian Detection," in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE. IEEE, 2021, pp. 9076–9082.

[33] Y. Li, L. Song, Y. Chen, Z. Li, X. Zhang, X. Wang, and J. Sun, "Learning Dynamic Routing for Semantic Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8553–8562.

[34] Z. Tian, T. He, C. Shen, and Y. Yan, "Decoders Matter for Semantic Segmentation: Data-Dependent Decoding Enables Flexible Feature Aggregation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3126–3135.

[35] Y.-H. Yang, T. E. Huang, S. R. Bulò, P. Kontschieder, and F. Yu, "Dense Prediction with Attentive Feature Aggregation," *ArXiv Preprint ArXiv:2111.00770*, 2021.

[36] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of Visual Saliency Detection with Comprehensive Information," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 2941–2959, 2018.

[37] Y. Liu, J. Han, Q. Zhang, and L. Wang, "Salient Object Detection via Two-Stage Graphs," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 4, pp. 1023–1037, 2018.

[38] Y. Zhou, T. Zhang, S. Huo, C. Hou, and S.-Y. Kung, "Adaptive Irregular Graph Construction-Based Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1569–1582, 2019.

[39] L. Wang, R. Chen, L. Zhu, H. Xie, and X. Li, "Deep Sub-Region Network for Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 728–741, 2020.

[40] L. Zhu, J. Chen, X. Hu, C.-W. Fu, X. Xu, J. Qin, and P.-A. Heng, "Aggregating Attentional Dilated Features for Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3358–3371, 2019.

[41] X. Hu, C.-W. Fu, L. Zhu, T. Wang, and P.-A. Heng, "SAC-Net: Spatial Attenuation Context for Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 1079–1090, 2020.

[42] R. Cong, Y. Zhang, L. Fang, J. Li, C. Zhang, Y. Zhao, and S. Kwong, "RRNet: Relational Reasoning Network with Parallel Multi-Scale Attention for Salient Object Detection in Optical Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–11, 2021.

[43] C. Fang, H. Tian, D. Zhang, Q. Zhang, J. Han, and J. Han, "Densely Nested Top-Down Flows for Salient Object Detection," *Science China Information Sciences*, pp. 1–13, 2021.

[44] D. Zhang, J. Han, Y. Zhang, and D. Xu, "Synthesizing Supervision for Learning Deep Saliency Network without Human Annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 7, pp. 1755–1769, 2019.

[45] Z. Liu, S. Shi, Q. Duan, W. Zhang, and P. Zhao, "Salient Object Detection for RGB-D Image by Single Stream Recurrent Convolution Neural Network," *Neurocomputing*, vol. 363, pp. 46–57, 2019.

[46] H. Wen, C. Yan, X. Zhou, R. Cong, Y. Sun, B. Zheng, J. Zhang, Y. Bao, and G. Ding, "Dynamic Selective Network for RGB-D Salient Object Detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 9179–9192, 2021.

[47] X. Zhu, Y. Li, H. Fu, X. Fan, Y. Shi, and J. Lei, "RGB-D Salient Object Detection via Cross-Modal Joint Feature Extraction and Low-Bound Fusion Loss," *Neurocomputing*, vol. 453, pp. 623–635, 2021.

[48] Q. Zhang, T. Xiao, N. Huang, D. Zhang, and J. Han, "Revisiting Feature Fusion for RGB-T Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1804–1818, 2020.

[49] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, and Q. Huang, "ASIF-Net: Attention Steered Interweave Fusion Network for RGB-D Salient Object Detection," *IEEE Transactions on Cybernetics*, vol. 51, no. 1, pp. 88–100, 2020.

[50] F. Wang, J. Pan, S. Xu, and J. Tang, "Learning Discriminative Cross-Modality Features for RGB-D Saliency Detection," *IEEE Transactions on Image Processing*, pp. 1–13, 2022.

[51] Z. Liu, W. Zhang, and P. Zhao, "A Cross-Modal Adaptive Gated Fusion Generative Adversarial Network for RGB-D Salient Object Detection," *Neurocomputing*, vol. 387, pp. 210–220, 2020.

[52] X. Shu, J. Yang, R. Yan, and Y. Song, "Expansion-Squeeze-Excitation Fusion Network for Elderly Activity Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–13, 2022.

[53] N. Liu, N. Zhang, and J. Han, "Learning Selective Self-Mutual Attention for RGB-D Saliency Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 756–13 765.

[54] B. Jiang, Z. Zhou, X. Wang, J. Tang, and B. Luo, "CmSalGAN: RGB-D Salient Object Detection with Cross-View Generative Adversarial Networks," *IEEE Transactions on Multimedia*, vol. 23, pp. 1343–1353, 2020.

[55] Z. Liu, K. Wang, H. Dong, and Y. Wang, "A Cross-Modal Edge-Guided Salient Object Detection for RGB-D Image," *Neurocomputing*, vol. 454, pp. 168–177, 2021.

[56] 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.

[57] Y. Zhao, J. Zhao, J. Li, and X. Chen, "RGB-D Salient Object Detection with Ubiquitous Target Awareness," *IEEE Transactions on Image Processing*, vol. 30, pp. 7717–7731, 2021.

[58] Z. Wu, G. Allibert, C. Stolz, C. Ma, and C. Demonceaux, "Modality-Guided Subnetwork for Salient Object Detection," in *2021 International Conference on 3D Vision (3DV)*, 2021, pp. 515–524.

[59] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking RGB-D Salient Object Detection: Models, Data Sets, and Large-Scale Benchmarks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2075–2089, 2020.

[60] X. Wang, S. Li, C. Chen, A. Hao, and H. Qin, "Depth Quality-Aware Selective Saliency Fusion for RGB-D Image Salient Object Detection," *Neurocomputing*, vol. 432, pp. 44–56, 2021.

[61] 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*. AAAI Press, 2021, pp. 1063–1071.

[62] F. Li, J. Zheng, Y.-f. Zhang, N. Liu, and W. Jia, "AMDFNet: Adaptive Multi-Level Deformable Fusion Network for RGB-D Saliency Detection," *Neurocomputing*, vol. 465, pp. 141–156, 2021.

[63] 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.

[64] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, "UC-Net: Uncertainty Inspired Rgb-D Saliency Detection via Conditional Variational Autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8582–8591.

[65] Y. Yang, Q. Qin, Y. Luo, Y. Liu, Q. Zhang, and J. Han, "Bi-Directional Progressive Guidance Network for RGB-D Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–15, 2022.

[66] C. Zhang, R. Cong, Q. Lin, L. Ma, F. Li, Y. Zhao, and S. Kwong, "Cross-Modality Discrepant Interaction Network for RGB-D Salient Object Detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2094–2102.

[67] R. Cong, J. Lei, H. Fu, W. Lin, Q. Huang, X. Cao, and C. Hou, "An Iterative Co-Saliency Framework for RGBD Images," *IEEE Transactions on Cybernetics*, vol. 49, no. 1, pp. 233–246, 2017.

[68] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou, "Saliency Detection for Stereoscopic Images Based on Depth Confidence Analysis and Multiple Cues Fusion," *IEEE Signal Processing Letters*, vol. 23, no. 6, pp. 819–823, 2016.

[69] Z. Chen, R. Cong, Q. Xu, and Q. Huang, "DPANet: Depth Potentiality-Aware Gated Attention Network for RGB-D Salient Object Detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 7012–7024, 2021.

[70] Y. Ma, D. Sun, Q. Meng, Z. Ding, and C. Li, "Learning Multiscale Deep Features and SVM Regressors for Adaptive RGB-T Saliency Detection," in *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 1, IEEE. IEEE, 2017, pp. 389–392.

[71] Z. Tu, T. Xia, C. Li, Y. Lu, and J. Tang, "M3S-NIR: Multi-Modal Multi-Scale Noise-Insensitive Ranking for RGB-T Saliency Detection," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval*, IEEE. IEEE, 2019, pp. 141–146.

[72] 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 *Chinese Conference on Image and Graphics Technologies*, Springer. Springer, 2018, pp. 359–369.

[73] J. Tang, D. Fan, X. Wang, Z. Tu, and C. Li, "RGBT Salient Object Detection: Benchmark and a Novel Cooperative Ranking Approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4421–4433, 2019.

[74] 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.

[75] W. Zhou, Q. Guo, J. Lei, L. Yu, and J.-N. Hwang, "ECFFNet: Effective and Consistent Feature Fusion Network for RGB-T Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–13, 2021.

[76] Z. Tu, Z. Li, C. Li, Y. Lang, and J. Tang, "Multi-Interactive Dual-Decoder for RGB-Thermal Salient Object Detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 5678–5691, 2021.

[77] W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, and W. Lin, "Unified Information Fusion Network for Multi-Modal RGB-D and RGB-T Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–16, 2021.

[78] J. Wang, K. Song, Y. Bao, L. Huang, and Y. Yan, "CGFNet: Cross-Guided Fusion Network for RGB-T Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[79] F. Huo, X. Zhu, L. Zhang, Q. Liu, and Y. Shu, "Efficient Context-Guided Stacked Refinement Network for RGB-T Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–14, 2021.

[80] G. Chen, F. Shao, X. Chai, H. Chen, Q. Jiang, X. Meng, and Y.-S. Ho, "CGMDRNet: Cross-Guided Modality Difference Reduction Network for RGB-T Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–17, 2022.

[81] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency Detection on Light Field," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2806–2813.

[82] N. Li, B. Sun, and J. Yu, "A Weighted Sparse Coding Framework for Saliency Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5216–5223.

[83] Y. Piao, Z. Rong, M. Zhang, and H. Lu, "Exploit and Replace: An Asymmetrical Two-Stream Architecture for Versatile Light Field Saliency Detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2020, pp. 11 865–11 873.

[84] T. Wang, Y. Piao, X. Li, L. Zhang, and H. Lu, "Deep Learning for Light Field Saliency Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8838–8848.

[85] M. Zhang, W. Ji, Y. Piao, J. Li, Y. Zhang, S. Xu, and H. Lu, "LFNet: Light Field Fusion Network for Salient Object Detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 6276–6287, 2020.

[86] M. Zhang, J. Li, W. Ji, Y. Piao, and H. Lu, "Memory-Oriented Decoder for Light Field Salient Object Detection," in *International Conference on Neural Information Processing Systems*, 2019, pp. 898–908.

[87] Y. Zhang, G. Chen, Q. Chen, Y. Sun, Y. Xia, O. Deforges, W. Hamidouche, and L. Zhang, "Learning Synergistic Attention for Light Field Salient Object Detection," in *Procedings of British Machine Vision Conference*, 2021, pp. 1–14.

[88] N. Liu, W. Zhao, D. Zhang, J. Han, and L. Shao, "Light Field Saliency Detection with Dual Local Graph Learning and Reciprocative Guidance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4712–4721.

[89] S. Yang, L. Zhang, J. Qi, H. Lu, S. Wang, and X. Zhang, "Learning Motion-Appearance Co-Attention for Zero-Shot Video Object Segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1564–1573.

[90] J. Li, W. Ji, Q. Bi, C. Yan, M. Zhang, Y. Piao, H. Lu *et al.*, "Joint Semantic Mining for Weakly Supervised RGB-D Salient Object Detection," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[91] J. Chen, Q. Li, H. Ling, D. Ren, and P. Duan, "Audiovisual Saliency Prediction via Deep Learning," *Neurocomputing*, vol. 428, pp. 248–258, 2021.

[92] D. Zhang, D. Meng, and J. Han, "Co-Saliency Detection via a Self-Paced Multiple-Instance Learning Framework," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 5, pp. 865–878, 2016.

[93] N. Zhang, J. Han, N. Liu, and L. Shao, "Summarize and Search: Learning Consensus-Aware Dynamic Convolution for Co-Saliency Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4167–4176.

[94] R. Cong, N. Yang, C. Li, H. Fu, Y. Zhao, Q. Huang, and S. Kwong, "Global-and-Local Collaborative Learning for Co-Salient Object Detection," *IEEE Transactions on Cybernetics*, pp. 1–11, 2022.

[95] S. Ren, Q. Wen, N. Zhao, G. Han, and S. He, "Unifying Global-Local Representations in Salient Object Detection with Transformer," *ArXiv Preprint ArXiv:2108.02759*, 2021.

[96] L. Tang, "CoSformer: Detecting Co-Salient Object with Transformers," *ArXiv Preprint ArXiv:2104.14729*, 2021.

[97] Z. Liu, Y. Wang, Z. Tu, Y. Xiao, and B. Tang, "TriTransNet: RGB-D Salient Object Detection with a Triplet Transformer Embedding Network," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4481–4490.

[98] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "TransCMD: Cross-Modal Decoder Equipped with Transformer for RGB-D Salient Object Detection," *ArXiv Preprint ArXiv:2112.02363*, 2021.

[99] Y. Wang, X. Jia, L. Zhang, Y. Li, J. Elder, and H. Lu, "Transformer-Based Network for RGB-D Saliency Detection," *ArXiv Preprint ArXiv:2112.00582*, 2021.

[100] X. Wang, B. Jiang, X. Wang, and B. Luo, "MutualFormer: Multi-Modality Representation Learning via Mutual Transformer," *ArXiv Preprint ArXiv:2112.01177*, 2021.

[101] 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.

[102] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-Token ViT: Training Vision Transformers From Scratch on ImageNet," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 558–567.

[103] Y. Mao, J. Zhang, Z. Wan, Y. Dai, A. Li, Y. Lv, X. Tian, D.-P. Fan, and N. Barnes, "Transformer Transforms Salient Object Detection and Camouflaged Object Detection," *ArXiv Preprint ArXiv:2104.10127*, 2021.

[104] J. Zhang, J. Xie, N. Barnes, and P. Li, "Learning Generative Vision Transformer with Energy-Based Latent Space for Saliency Prediction," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1–16, 2021.

[105] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[106] Q. Hou, D. Zhou, and J. Feng, "Coordinate Attention for Efficient Mobile Network Design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 713–13 722.

[107] Z. Shen, M. Zhang, H. Zhao, S. Yi, and H. Li, "Efficient Attention: Attention with Linear Complexities," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3531–3539.

[108] 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. Springer, 2020, pp. 213–229.

[109] 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. Springer, 2014, pp. 92–109.

[110] 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. IEEE, 2014, pp. 1115–1119.

[111] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging Stereopsis for Saliency Analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE. IEEE, 2012, pp. 454–461.

[112] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth Enhanced Saliency Detection Method," in *Proceedings of International Conference on Internet Multimedia Computing and Service*, 2014, pp. 23–27.

[113] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-Induced Multi-Scale Recurrent Attention Network for Saliency Detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7254–7263.

[114] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "RGBT Salient Object Detection: A Large-Scale Dataset and Benchmark," *IEEE Transactions on Multimedia*, pp. 1–14, 2022.

[115] J. Zhang, M. Wang, L. Lin, X. Yang, J. Gao, and Y. Rui, "Saliency Detection on Light Field: A Multi-Cue Approach," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 3, pp. 1–22, 2017.

[116] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient Object Detection: A Benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[117] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-Measure: A New Way to Evaluate Foreground Maps," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4548–4557.

[118] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-Tuned Salient Region Detection," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE. IEEE, 2009, pp. 1597–1604.

[119] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-Alignment Measure for Binary Foreground Map Evaluation," in *International Joint Conferences on Artificial Intelligence Organization*, 2018, pp. 698–704.

[120] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency Filters: Contrast Based Filtering for Salient Region Detection," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE. IEEE, 2012, pp. 733–740.

[121] J. Wei, S. Wang, and Q. Huang, "F^3Net: Fusion, Feedback and Focus for Salient Object Detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2020, pp. 12 321–12 328.

[122] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese Network for RGB-D Salient Object Detection and Beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2021.

[123] W.-D. Jin, J. Xu, Q. Han, Y. Zhang, and M.-M. Cheng, "CDNet: Complementary Depth Network for RGB-D Salient Object Detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3376–3390, 2021.

[124] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical Alternate Interaction Network for RGB-D Salient Object Detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3528–3542, 2021.

[125] 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*, pp. 1–14, 2021.

[126] N. Huang, Y. Yang, D. Zhang, Q. Zhang, and J. Han, "Employing Bilinear Fusion and Saliency Prior Information for RGB-D Salient Object Detection," *IEEE Transactions on Multimedia*, pp. 1–14, 2021.

[127] 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.

[128] Y. Piao, Z. Rong, M. Zhang, X. Li, and H. Lu, "Deep Light-Field-Driven Saliency Detection From a Single View," in *International Joint Conference on Artificial Intelligence*, 2019, pp. 904–911.

[129] Y. Piao, Y. Jiang, M. Zhang, J. Wang, and H. Lu, "PANet: Patch-Aware Network for Light Field Salient Object Detection," *IEEE Transactions on Cybernetics*, pp. 1–13, 2021.

[130] Q. Zhang, S. Wang, X. Wang, Z. Sun, S. Kwong, and J. Jiang, "A Multi-Task Collaborative Network for Light Field Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1849–1861, 2020.

[131] D. Jing, S. Zhang, R. Cong, and Y. Lin, "Occlusion-Aware Bi-Directional Guided Network for Light Field Salient Object Detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1692–1701.

[132] 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.

[133] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," in *Advances in Neural Information Processing Systems*, 2021, pp. 1–14.
[134] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

Bin Tang is a lecturer in School of Artificial Intelligence and Big Data, Hefei University, China. He received his Ph.D. from Fudan University, China in 2008. His research interests include computer vision.

Zhengyi Liu is a professor in School of Computer Science and Technology, Anhui University, China. She received her B.S., M.S., and Ph.D. from Anhui University, China in 2001, 2004 and 2007, respectively. Her research interests include image and video processing, computer vision.

Yacheng Tan is a M.S. Candidate of Anhui University. He received his B.S. from Anhui Polytechnic University, China in 2020. His research interests include image and video processing, computer vision.

Qian He is a M.S. Candidate of Anhui University. She received her B.S. from Jianghuai College of Anhui University, China in 2021. Her research interests include image and video processing, computer vision.