Neural Computing and Applications (2023) 35:17733–17747
https://doi.org/10.1007/s00521-023-08640-8

ORIGINAL ARTICLE

Transformer guidance dual-stream network for salient object detection in optical remote sensing images

Yi Zhang[1] · Jichang Guo[1] · Huihui Yue[1] · Xiangjun Yin[1] · Sida Zheng[1]

Received: 20 October 2022 / Accepted: 2 May 2023 / Published online: 20 May 2023

Abstract

Salient object detection (SOD) has achieved remarkable performance in natural scene images (NSIs). However, current SOD methods still face serious challenges in processing optical remote sensing images (RSIs) due to cluttered backgrounds, diverse scales, and different views, which are distinguished from NSIs. In this paper, a transformer guidance dual-stream network (TGDNet) is proposed for SOD in optical RSIs. The key insight is to extract multi-scale features by global receptive fields and separately refine them according to the characteristics of feature hierarchies. Specifically, inspired by the long-range dependencies of transformer, a transformer guidance dual-stream strategy is proposed to compensate the extracted details such as boundaries and edges using global information. To overcome the issue of diverse scales of salient objects in optical RSIs, a sequence inheritance channel attention module is built to focus more on high-level semantic features at different scales. In addition, a pyramid spatial attention module is elaborately designed to refine low-level features as well as to suppress background interference for accurate SOD in optical RSIs. At last, a coarse-to-fine decoder is utilized to progressively predict salient objects. In the experiment, the EORSSD dataset is employed to train and evaluate the proposed TGDNet. It achieves performance of 0.0049, 0.8964, and 0.9286 in terms of MAE, F-measure, and S-measure, respectively. Furthermore, ORSSD dataset is also utilized to evaluate the generality. Experimental results demonstrate the advantages of TGDNet over the state-of-the-art SOD methods.

Keywords Salient object detection · Transformer · Attention mechanism · Optical remote sensing images

1 Introduction

Salient object detection (SOD) aims to locate the most visually attractive objects or regions in a scene [1, 2]. In the past few years, SOD has attracted increasing attention in computer vision community and shown the efficacy in visual tracking [3], semantic segmentation [4, 5], image captioning [6, 7], manipulation [8], and image retrieval [9]. Optical remote sensing images (RSIs) are widely used in many fields, such as agriculture and military [10, 11], showing the promising applications. As one of the typical computer vision tasks, SOD in optical RSIs appeals to increasing research interests in recent years. Similar to natural scene images (NSIs), SOD in optical RSIs is decomposed into detecting and segmenting regions and objects of interest. However, due to the enormous disparity between NSIs and optical RSIs, SOD methods for NSIs are hardly directly applied in optical RSIs. As shown in Fig. 1, optical RSIs have three prominent characteristics: (1) since optical RSIs are acquired automatically by remote sensors at high altitudes, the scale of salient objects varies greatly; (2) optical RSIs are vertical views collected from an overhead perspective, containing objects with various rotational orientations; (3) optical RSIs have more complicated background patterns, more illumination variation,

✉ Jichang Guo
jcguo@tju.edu.cn

Yi Zhang
zhangyi123@tju.edu.cn

Huihui Yue
yuehuihui@tju.edu.cn

Xiangjun Yin
yinxiangjun@tju.edu.cn

Sida Zheng
zhengsida@tju.edu.cn

[1] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

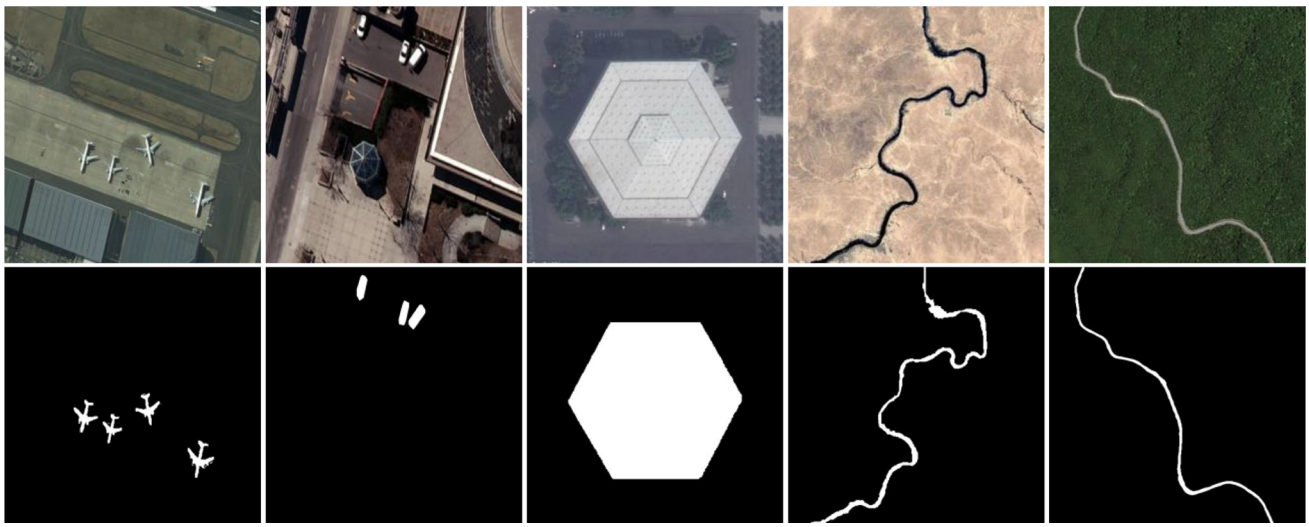

Fig. 1 Samples of optical RSIs. Top row shows the optical RSIs, and bottom row shows the corresponding salient regions

more diverse texture structures, and more fragmented distribution. Therefore, SOD for optical RSIs faces huge challenges in achieving accurate detection: (1) the need for larger receptive fields due to the scale diversity; (2) the conflict between achieving large receptive fields and remaining rich local details.

For the first challenge, many existing CNN-based methods are dedicated to expand receptive fields [12, 13]. They mainly benefit from increasing network depth, designing effective fusion strategies across layer hierarchies, and constructing specific convolutional operations. However, few of them essentially consider the fact that the limited receptive field of convolution still constrain the performance of CNNs. Besides, the actual receptive field is much smaller than the theoretical one [14, 15]. Recent flourishing transformers are well suited for this challenge. Transformers exploit multi-head self-attention (MHSA) to model long-term dependencies [16, 17]. Unlike convolution that keeps a limited receptive field, the MHSA of transformers has dynamic weights and a global receptive field. Inspired by this, we make the first attempt to exploit transformer to SOD in optical RSIs.

For the second challenge, the trade-off between the size of receptive fields and the richness of local details is to be weighed. A larger receptive field is attained at the expense of details, and vice versa. Existing efforts mainly focus on developing sophisticated decoder structures [18, 19] and effective interaction strategies between high-level and low-level information [20–22] to achieve better balance. However, they usually ignore the decisive role played by the encoder. Considering the outstanding capability of convolution in preserving details, we make full use of such a characteristic to upgrade the encoder, which helps to obtain the feature representations with both global information and good local details.

To deal with the above-mentioned two challenges, we propose a novel transformer guidance dual-stream network (TGDNet). Firstly, to generate salient predictions with uniformly highlighted regions, we develop a transformer guidance dual-stream strategy for encoding. It focuses on the complementarity between the global features from transformer and the detail information from convolution. Specifically, the transformer guidance dual-stream encoder (TGDE) consists of a transformer stream and a local aggregation stream. Global contextual information is efficiently modeled through the transformer stream, and then is integrated into the local aggregation stream to provide global knowledge guidance on local details. Secondly, to extract the salient objects of different scales, we propose a sequence inheritance channel attention module (SICAM), which adaptively extracts multi-scale information from high-level features. Thirdly, a pyramid spatial attention module (PSAM) is designed to eliminate redundant information from low-level features and suppress background interference. Our contributions are summarized as follows:

- We propose a transformer guidance dual-stream network (TGDNet) for SOD in optical RSIs. The proposed transformer guidance dual-stream strategy endows accurate global feature representation by explicitly aggregating global information from transformer with local details from convolution. Such designs enable our method to detect complete and sharp salient objects in optical RSIs.
- A novel sequence inheritance channel attention module (SICAM) is proposed to extract task-specific multi-scale high-level features. Meanwhile, a pyramid spatial attention modul (PSAM) is deployed to refine the detail

information of low-level features. The proposed SICAM coupled with the PSAM effectively extracts high-level and low-level features that boost the SOD performance.
- Extensive experiments on two benchmark datasets validate the effectiveness of our proposed TGDNet and also demonstrate that our method outperforms 17 state-of-the-art methods.

The rest of this paper is organized as follows. Related works to SOD and vision transformer are introduced in Sect. 2. In Sect. 3, we detail each component of our method. The results of ablation study and comparison experiments are shown in Sect. 4. Finally, the conclusion is drawn in Sect. 5.

2 Related work

In this section, we introduce the existing SOD methods in NSIs and optical RSIs. Besides, we briefly introduce the recent progress of transformer in computer vision.

2.1 Salient object detection for NSIs

Benefiting from the rapid development of deep learning, SOD in NSIs [18, 20, 23] has made a significant breakthrough in recent years, especially after the FCN-based method [24] was proposed. For example, Hou et al. [25] used the short connection on the foundation of holistically-nested edge detection [26]. The high-level features are integrated with low-level features so that the semantic information in deep layers can guide the detailed information in shallow layers. Liu et al. [27] applied the U-shape architecture to SOD, allowing the attention mechanism to selectively focus on global contextual information at high levels and integrating local information at low levels at the same time. Zhao et al. [20] integrated the high-level features and low-level features to investigate the complementarity between salient edge information and salient object information. Qin et al. [28] focused on boundary quality and presented a residual refinement module to further refine the edge of the salient map after an encoder-decoder network. Su et al. [23] designed a boundary localization stream and an interior perception stream to detect the object interiors and boundaries, respectively, producing uniformly highlighted interiors and clear edges. Liu et al. [21] deployed a global guidance module and a feature aggregation module on the top-down path. Zhao et al. [29] utilized the spatial attention mechanism in shallow layers and channel attention mechanism in deep layers to achieve subtle attention on the salient boundary and salient semantic regions, respectively. Pang et al. [18] proposed the aggregate interaction modules, which integrates the features from adjacent levels. Siris et al. [30] proposed a context-aware learning approach to explicitly learn and enhance the contextual relationships between salient objects and scene contexts. Wu et al. [31] decomposed salient objects into edge, skeleton, and saliency maps, then designed a completion network, utilizing the obtained edge and skeleton maps to refine saliency maps. Nevertheless, due to the gap between optical RSIs and NSIs, directly adopting SOD methods for NSIs to optical RSIs is unlikely to yield effective SOD.

2.2 Salient object detection for optical RSIs

Despite the significant progress of SOD in NSIs, SOD in optical RSIs still falls behind. Zhao et al. [32] proposed a sparsity-guided saliency detection model for optical RSIs, which uses a sparse representation to acquire the high-level global information and background cues. Zhang et al. [33] developed a self-adaptively multiple feature fusion model to take advantage of the intrinsic relationship among different cues. Later, Li et al. [34] introduced deep learning into SOD in optical RSIs and constructed an end-to-end deep network. Li et al. [35] designed a parallel down-up fusion network, taking full care of the in-path low-level and high-level features and cross-path multi-resolution features. Zhang et al. [12] proposed a dense attention fluid network to adaptively capture long-range semantic context relationships and to generate high-level attention maps by introducing shallow attention into deep layers. Zhou et al. [36] adopted an edge-aware multi-scale feature integration network that emphasizes the boundary by applying edge information of salient objects. Tu et al. [37] designed a joint boundary and region learning scheme based on a bidirectional feature transformation to optimize both features. For multiple features that affect optical RSIs, Li et al. [38] exploited attraction mechanism to achieve content complementarity among these features, which highlights the salient objects at different scales. Cong et al. [39] proposed a relational reasoning module for high-level features to extract semantic information and designed a parallel multi-scale attention module for low-level features to efficiently recover details. Tu et al. [37] proposed a multi-scale joint boundary and region model to obtain robust multi-scale region features by simultaneously embedding the boundary features. These methods strive to extract multi-scale features of optical RSIs, but barely escape from the limited receptive field of convolution.

2.3 Transformer in computer vision

Transformer is a self-attention mechanism that dominates in Natural Language Processing (NLP) because of its ability of

capturing long-range dependencies among sequence elements. Its outstanding performance of has dramatically attracted the enthusiasm of researchers and been introduced into many fields of computer vision, such as object detection [40, 41], semantic segmentation [42–44], panoramic segmentation [45], salient object detection [14, 15], and image generation [46]. As a pioneering work, Dosovitskiy et al. [41] proposed the vision transformer (ViT) and applied it to image classification. Many downstream tasks applied ViT as the backbone of their models. Liu et al. [47] proposed a Swin Transformer, which limits self-attention to non-overlapping local windows but allows cross-window connection by shifted windows for effectively extracting features with a hierarchical structure. Carion et al. [40] proposed an end-to-end object detection with transformers (DETR), regarding object detection as a direct set prediction problem and modeling the relationship between them. Zheng et al. [44] treated semantic segmentation as a sequence-to-sequence prediction task, where transformer is employed as the encoder that learns more semantic information under large receptive fields by transform the image into a sequence of patches without downsampling. Wang et al. [45] proposed the first end-to-end panoramic segmentation model with a mask transformer to predict class-labeled masks. Liu et al. [15] designed a visual saliency transformer for RGB-D SOD in NSIs. Observing that the transformer backbone can provide accurate structure modeling, Mao et al. [14] adopted a dense transformer backbone for fully supervised and weakly supervised salient object detection in optical NSIs. From a convolution-free sequence-to-sequence perspective, Liu et al. [48] developed a pure transformer architecture model for SOD, in which a token upsampling method is proposed to get high-resolution detection results. In SOD, transformer methods capture salient features under the global view and can thus give a satisfactory prediction on the location of salient objects. However, since the detailed information is underutilized in transformer methods, rough edges and unevenly highlighted salient regions are typical when transformer-only backbone SOD methods for NSIs are applied to optical RSIs. In this study, we exploit the large receptive field of transformer and the small one of convolution to sufficiently extract and effectively integrate global and local features that could produce the final high-quality salient results.

3 Proposed method

3.1 Motivation

As described in Sect. 1, the key challenge of SOD for optical RSIs lies in balancing the need for accommodating large receptive fields to encompass scale diversity with retaining local information. Transformers leverage MHSA to establish long-term dependencies, while convolutions excel in extracting details. Inspired by this, we propose a transformer guidance dual-stream strategy that complements global features and detail information to boost SOD performance, as detailed in Sect. 3.3. For the issue of scale diversity and redundant information, we consider that processing features separately at high and low levels facilitate further extraction and refinement. To this end, we design the SICAM and PSAM modules using channel attention and spatial attention mechanisms, respectively, which we will elaborate on in Sect. 3.4 and Sect. 3.5.

3.2 Overview

The proposed TGDNet is mainly composed of four components, a transformer guidance dual-stream encoder (TGDE), two sequence inheritance channel attention modules (SICAMs), a pyramid spatial attention module (PSAM), and a decoder. The overall architecture is shown in Fig. 2.

Specifically, an input optical RSI is fed into both the transformer stream and the local aggregation stream of TGDE. Transformer stream extracts global features, then these features are forwarded to the semantic guidance module (SGM) to be combined with the detailed features from local aggregation stream. In this way, the complementary feature representations are achieved in the SGM. Note that, the SGM also integrates the deepest features from transformer stream and local aggregation stream to further extract semantic information on the top of encoder. Then, the SICAM and PSAM eliminate the redundant information and reassign feature weights to ensure that the features are refined and highly related to salient regions. The semantic information is enhanced by the SICAM that receives features from the 4th and 5th layers of encoder. In contrast, the detail information is distinguished by the PSAM for the remaining shallow layer features. Subsequently, the decoder consists of five convolutional blocks, each block fuses the features from its previous block, its corresponding encoder stage, and the deep layer guidance features from the SGM. Thus, the salieny maps are progressively refined from top to down with the supervision. At last, the salient prediction is obtained with precise edges and complete structure. In what follows, we will detail each component.

3.3 Transformer guidance dual-stream strategy

For any two pixels of a salient object in an image, the learned features should be consistent theoretically. But the feature extracted by convolution is essentially local, which leads to the corrupted consistency of the feature when

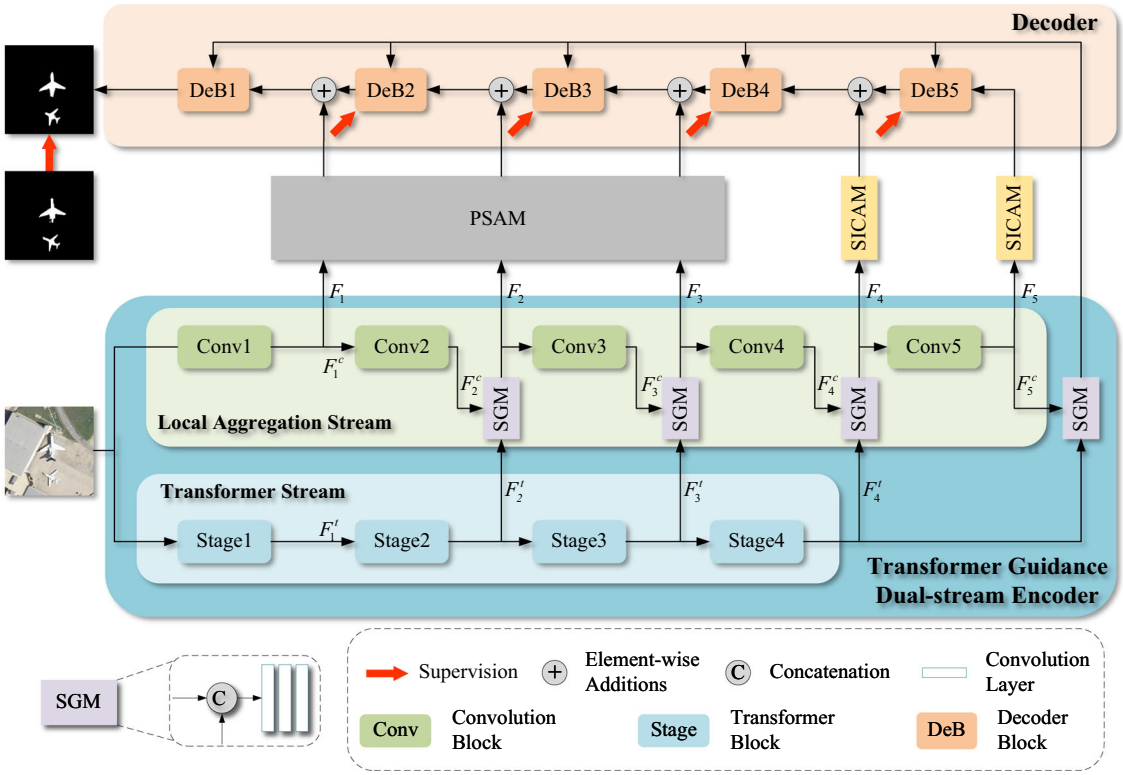

Fig. 2 Overview of the proposed TGDNet. TGDNet consists of four main components: Transformer Gudiance Dual-stream Encoder (TGDE), Sequence Inheritance Channel Attention Module (SICAM), Pyramid Spatial Attention Module (PSAM), and Decoder. TGDE consists of 1) Transformer Stream extracts global features $\{F_i^t, i = 1, 2, 3, 4\}$; 2) Local Aggregation Stream captures local features $\{F_i^c, i = 1, 2, 3, 4, 5\}$, and four simple semantic guidance modules (SGMs) integrate the global and local features as the encoder feature representations $\{F_i, i = 1, 2, 3, 4, 5\}$. SICAM and PSAM exploit multi-scale semantic features and refine detail information for high-level and low-level features, respectively. After the progressive incorporation by decoder on the top-down path and the supervision of each decoder block, an accurate salienct map of the input image can be achieved

internal features of the object are discontinuous and the background is diverse. To alleviate this problem, a commonly used method is to continuously downsample the feature map so that its scale approaches that of the convolution receptive field. However, such a method may still lead to inconsistent features for large-scale salient objects or large-size input images, where the features obtained on the deepest remain local and cannot cover the complete salient object. In addition, successive downsampling could be severely and irreversibly destructive to the information of the salient object with small scales, raising the difficulty of detecting such objects. In optical RSIs, such issues become more severe due to the diverse scales of salient objects. To cope with this challenging issue, we propose a transformer guidance dual-stream strategy. A transformer guidance dual-stream encoder exploits the long-range dependencies to preserve useful information of small-scale salient objects while maintaining the feature consistency of large-scale ones.

The transformer guidance dual-stream encoder consists of a transformer stream and a local aggregation stream, the former is implemented by Swin-S [47], and the latter is constructed based on the commonly ResNet-50, in which we removed the last average pool and fully connected layers. In addition, the input of the convolutional block in local aggregation stream, $Conv\ i, i = 3, 4, 5$, derives from the SGM rather than from the previous convolutional block $Conv\ i - 1$.

For an input image $I \in \mathbb{R}^{3\times H\times W}$, transformer stream generates four level features from shallow to deep as $F^t = \{F_i^t, i = 1, 2, 3, 4\}$. Similarly, five scale features are captured by the local aggregation stream as $F^c = \{F_i^c, i = 1, 2, 3, 4, 5\}$. We integrate F_i^t with the same level of F_i^c by a simple SGM, as shown in Fig. 2, which incorporates the perception of the global context to guide the next convolution block. This process can be formulated as:

$$F_i = \begin{cases} Conv(Cat(F_i^t, F_i^c); W_i), & i = 2, 3, 4 \\ F_i^c, & i = 1, 5 \end{cases} \quad (1)$$

where each convolution operator $Conv$ is followed by a batch normalization (BN) and a ReLU, W denotes trainable parameters, and Cat indicates the channel concatenation.

Considering the large size of the feature $(C \times \frac{H}{4} \times \frac{W}{4})$ extracted via *Stage*1 of the transformer stream would undoubtedly increase the computational complexity of SGM, and the feature still lacks information with actual values for global guidance, we discard the exploitation of F_1^t to the local aggregation stream. When $i = 2, 3, 4$, F_i^t and F_i^c are incorporated into SGM to obtain F_i, resulting in the global information of F_i^t selectively flowing into local aggregation stream. Under the guidance of global information, local aggregation stream discards the way of indiscriminate extraction on features, thereby suppressing the interference of local non-salient features. That is, local aggregation stream could extract local information under implicitly delineating the target domain. Moreover, the deepest features (F_4^t and F_5^c) of both streams are also integrated into SGM to further extract the global semantic information, and then the obtained deep features are fed into each decoder block as a reference during decoding. Figure 3 suggests the effectiveness of our transformer guidance dual-stream strategy, where transformer and convolution are superior in global and detail extraction, respectively, and SGM preserves both advantages to produce feature maps by aggregating global and detail information together.

3.4 Sequence inheritance channel attention module

For the diverse scales of optical RSIs, Zhou et al. [36] and Li et al. [34] downsampled the input image with different degrees to extract multi-scale information in parallel paths. Zhang et al. [12] employed dense attention to bridge the information between high-levels and low-levels. But these methods only take into account the interactions of the scale information between "inter-features," ignoring the potential multi-scale features in "intra-feature." In addition, channel attention is regarded as a commonly used mechanism for refining deep features [29, 49]. However, when the current channel attention mechanism is applied to the hierarchical feature maps, the generated attention map is only responsible for its source feature map, but fails to reflect the joint focus of the series of feature maps. This is due to the transfer of high-response channels among hierarchical features caused by the convolution procedure, as shown in Fig. 4a. Because the attention map is dominated by a single feature map, the incidence of the channel attention mechanism gets weakened, and the risk of attending to the non-focus is increased. Therefore, we propose a SICAM as shown in Fig. 5, which extracts multi-scale hierarchical information in "intra-feature" and assigns joint channel weights to them.

Specifically, in the process of multi-scale feature extraction, as shown in Fig. 4b, the attributes of the elements within the channel sequence are guaranteed to remain independent of each other by depth-wise separable convolution, avoiding the transfer of high-response channels owing to channel communication. Therefore, the element response attributes of each feature are fully inherited from the corresponding elements of its previous map. For high-level features $F_i, i = 4, 5$, the hierarchical "intra-features" are generated by:

$$F_{i1}^m = DepthConv(F_i; W_{i1}), \tag{2}$$

$$F_{i2}^m = DepthConv(F_{i1}^m; W_{i2}), \tag{3}$$

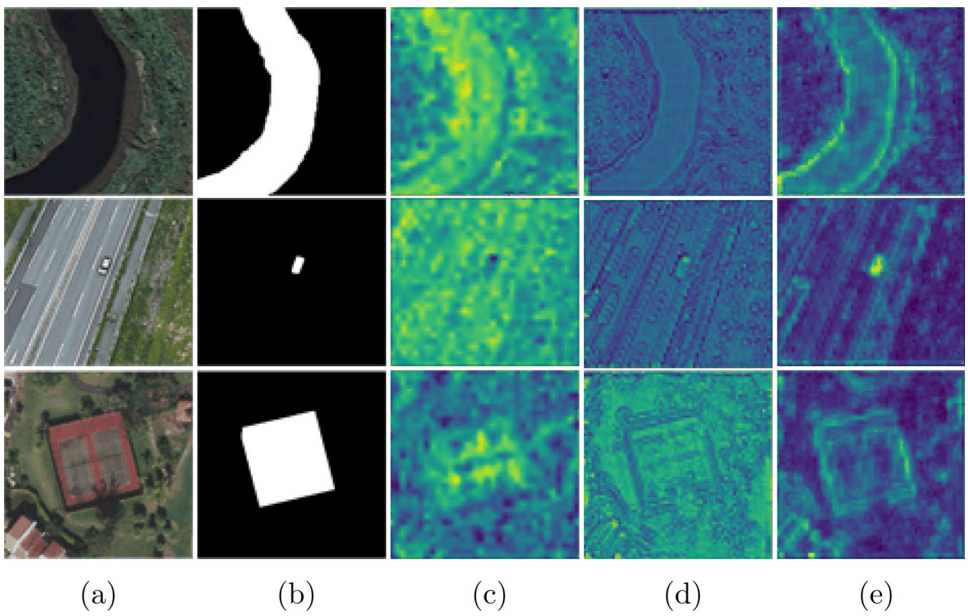

Fig. 3 Visualizing feature maps of SGM. **a** Input image; **b** GT; **c–e** show the visualizations of feature maps F_2^t, F_2^c, and F_2, respectively. After SGM, the global semantic information from transformer and the detailed information from convolution are successfully integrated

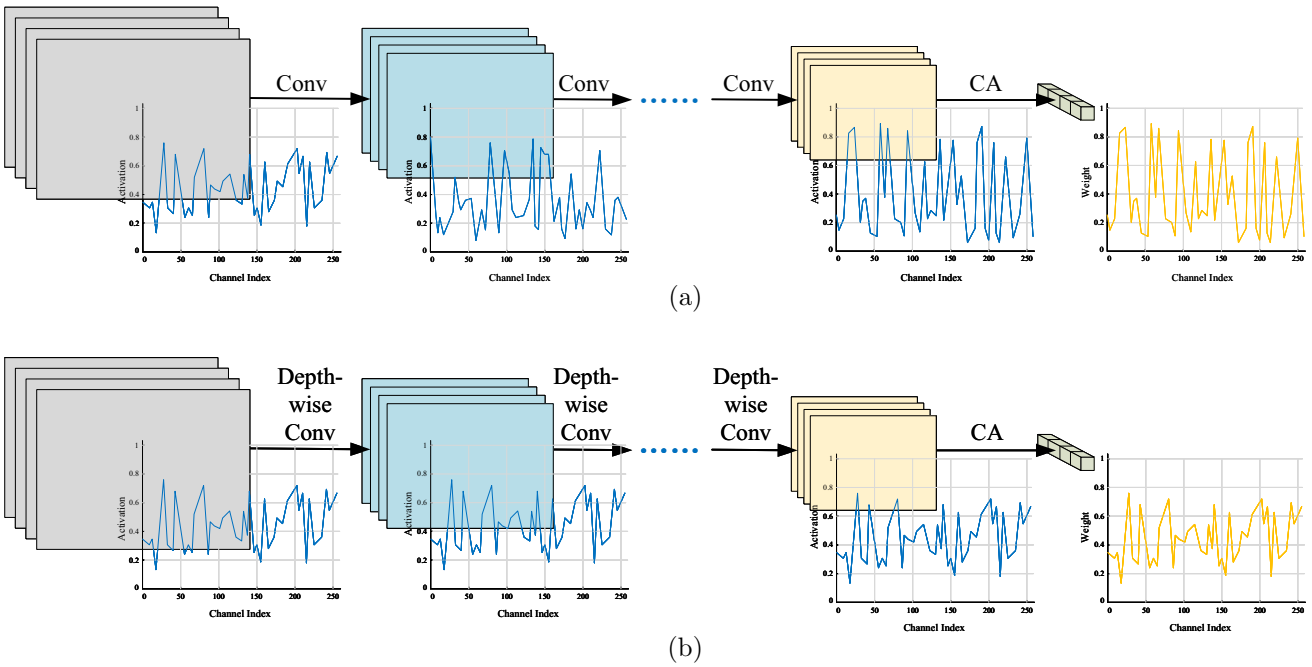

Fig. 4 Two kinds of methods of computing attention weights. **a** Traditional multi-scale feature and channel attention weight extraction. **b** Multi-scale feature and channel attention of the proposed SICAM. The coordinates with the blue line represent the activation of different channels for salient objects, and the coordinates with the gold line represent the response weights of different channels for salient objects. It can be seen that the response weight line is only matched to the activation line of final feature map in (**a**), while the response weight line keeps consistent with the activation line of all its previous features in (**b**). The depth-wise separable convolution keeps the independence between feature channels, guaranteeing no transfer of channel responses. Thus the attention weights generated by SICAM could match the feature maps of any level

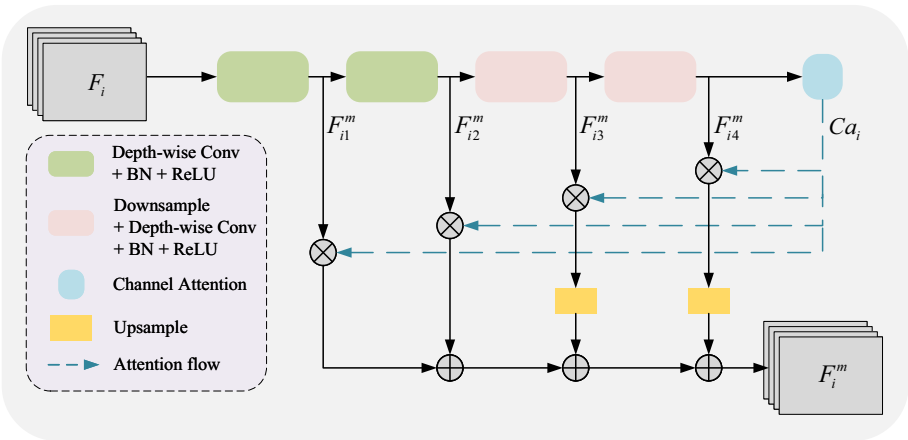

Fig. 5 Illustration of the proposed SICAM. The feature map is first fed to the stacked depth-wise separable convolution to extract multi-scale features without disrupting the order of feature channel sequence, after which the channel weights obtained are assigned to each multi-scale feature map. Then the feature maps are interpolated to the same size and summed to generate the refined features

$$F^m_{i3} = DepthConv(Down(F^m_{i2}); W_{i3}), \tag{4}$$

$$F^m_{i4} = DepthConv(Down(F^m_{i3}); W_{i4}), \tag{5}$$

where *DepthConv* denotes a depth-wise separable convolution with a kernel size of 3×3, followed by a BN and a ReLU. *Down* represents the downsample pooling with the scales of 2. $F^m_{ij}, j = 1, 2, 3, 4$ denote the different scale features corresponding to F_i. For the minimized scale feature F^m_{i4}, we squeeze it to $d \times 1 \times 1$, where d represents the channel number of the ith-level encoder feature. Then, two consecutive fully connected layers and a Sigmoid operation are followed to generate the attention weights:

$$Ca_i = \sigma(fc2(relu(fc1(Pool(F^m_{i4}); W_{i5\_1})); W_{i5\_2})), \tag{6}$$

where Ca_i is denoted as the channel response weights corresponding to the ith-encoder level. *Pool* means global average pooling, and σ represents Sigmoid operation. W_{i5_1} and W_{i5_2} denote the trainable weights of fully connected layers. Subsequently, the channel attention is performed on the generated multi-scale features mentioned

above. We perform element-wise multiplication of Ca_i with each F_{ij}^m, upsample the feature maps to the same size and add up them by:

$$F_i^m = Ca_i * F_{i1}^m + Ca_i * F_{i2}^m + Up2(Ca_i * F_{i3}^m) + Up4(Ca_i * F_{i4}^m), \tag{7}$$

where $Up2$ and $Up4$ represent upsample with scale of 2 and 4, respectively. F_i^m is the salient feature map after refinement of SICAM, which will be sent to the ith-level of decoder.

3.5 Pyramid spatial attention module

For the complicated background of optical RSIs, we propose a pyramid spatial attention module (PSAM) to refine low-level features. It alleviates the obstruction of background noise and highlights foreground information, producing more distinguishable feature representations with sharp edges. As shown in Fig. 6, this module first progressively refines the inputs from deeper to shallower with a stacked pyramid structure. With the involvement of semantic information from deeper layers, the redundant features of shallower layers are suppressed. The process can be expressed as:

$$F_3^s = conv(F_3; W_3), \tag{8}$$

$$F_{2\_1}^s = conv(Cat(F_2, Up2(F_3^s)); W_{2\_1}), \tag{9}$$

$$F_2^s = conv(F_{2\_1}^s; W_{2\_2}), \tag{10}$$

$$F_1^s = conv(conv(Cat(conv(Cat(F_1, Up2(F_{2\_1}^s)); W_{1\_1}), Up2(F_2^s)); W_{1\_2}); W_{1\_3}), \tag{11}$$

where $F_i^s, i = 1, 2, 3$ are the refined feature maps in the ith-layer. Then, the spatial attention mechanism [50] is applied on them. Moreover, the generated spatial attention maps are also cascaded from deeper to shallower so as to make them more representative. Lastly, the attention maps generated by each layer are multiplied with corresponding encoder features to obtain the refined feature maps:

$$F_3^m = Sa(F_3^s; W_{m3}) * F_3, \tag{12}$$

$$F_2^m = conv\_7(Cat(Up2(F_3^m), Sa(F_2^s)); W_{m2})) * F_2, \tag{13}$$

$$F_1^m = conv\_7(Cat(Up2(F_2^m), Sa(F_1^s)); W_{m1})) * F_1, \tag{14}$$

where $conv\_7$ denotes the convolution operation with a 7×7 kernel followed by a BN and a ReLU. Sa represents the spatial attention module, which is achieved by concatenating the outputs of a global average pooling and a global max pooling, followed by a 7×7 convolution and a Sigmoid layer.

3.6 Loss

The loss is defined as the sum of the losses of all outputs:

$$L = \sum_{k=1}^{K} \alpha_k l_k, \tag{15}$$

where $K = 5$ indicates the total number of outputs, α_k denotes the weight of each loss, and l_k is the loss at the output of the k-th block of decoder. Similar to most SOD experiments, we choose the binary cross entropy (BCE) loss as the loss function, which is formulated as:

$$l_k = -(y \log(p_k) + (1 - y) \log(1 - p_k)), \tag{16}$$

where p_k and y represent the salient prediction and the ground truth (GT), respectively.

4 Experiments

In this section, we first introduce the experimental settings, including the benchmark datasets, the evaluation metrics, and the implementation details. Then, we perform a series

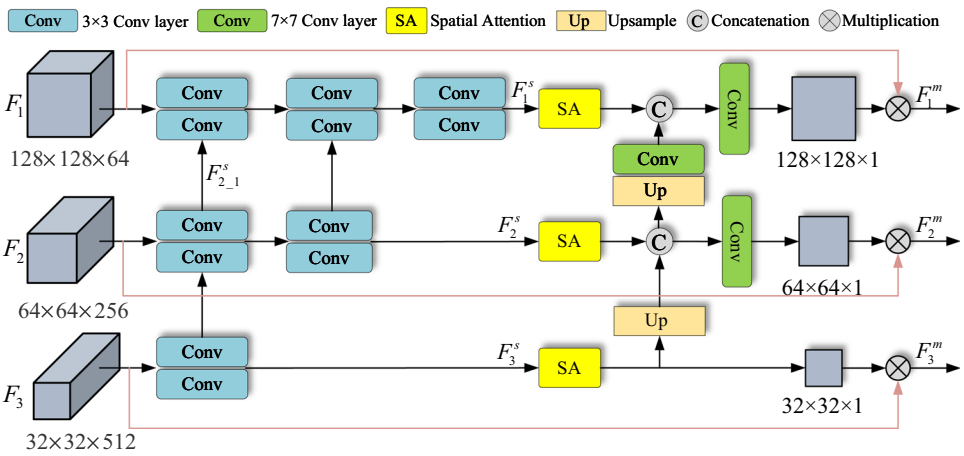

Fig. 6 Illustration of the proposed PSAM. The features F_1, F_2, and F_3 from encoder are first progressively elaborated through a pyramid structure, followed by the spatial attention operation. Then the attention maps obtained are further detailed from deeper to shallower, eventually multiplied with corresponding inputs to obtain the refined results

of ablation studies to verify the effectiveness of each module of the proposed method. Finally, we compare the proposed approach with state-of-the-art methods.

4.1 Experiment setups

4.1.1 Datasets

We conduct experiments on two benchmark optical RSI datasets: ORSSD[1] and EORSSD.[2] ORSSD consists of 600 training images and 200 testing images. EORSSD contains 1400 pairs of images for training and 600 images for testing. Both datasets contain cluttered backgrounds, objects with wide-scale variation, and rich object types, so they could fairly represent the properties of optical RSIs. We train our network on EORSSD dataset and test on the testing subset of ORSSD and EORSSD.

4.1.2 Evaluation metrics

We utilize four measures to evaluate the performance of our method: precision-recall (PR) curves, F-measure score, S-measure score, and mean absolute error (MAE). *PR curve* is a standard metric for evaluating the predicted saliency probability maps [51]. Precision score and recall score are both obtained by comparing the binary salient map under the threshold from 0 to 255 with GT. F-measure [52] is denoted as F_β, which is a comprehensive evaluation measure calculated by the weighted harmonic mean of precision and recall scores:

$$F_\beta = \frac{(1+\beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}, \tag{17}$$

where β^2 is set to 0.3 to weight precision more than recall. A larger F-measure means greater effectiveness of method. The maximum F_β ($\max F_\beta$) is reported in this paper. MAE [18, 39] directly compares the difference between the salient map S and the paired ground truth G, which can be denoted as:

$$\text{MAE} = \frac{1}{W \times H} \sum_{r=1}^{W} \sum_{c=1}^{H} |S(r,c) - G(r,c)|, \tag{18}$$

where W and H represent the width and height of image and (r, c) denotes the pixel coordinates. The superior performance of the method is reflected in the small value of MAE. S-measure [36, 53] compares the structural similarity of the salient map and GT. The larger value of the S-measure means the better performance. It is defined as:

$$S = \alpha \times S_o + (1-\alpha) \times S_r, \tag{19}$$

where α is set to 0.5 as suggested in [53], S_o and S_r denote the object similarity and the region similarity, respectively.

4.1.3 Implementation details

The proposed TGDNet is implemented by PyTorch with an NVIDIA GeForce RTX 3090 GPU. During training, each input image is resized to 256×256, and data augmentation is utilized to improve the generalizability and robustness of the model. We employ the Adamw [54] optimizer to train our network with a weights decay of 0.01 and an initial rate of 1e-6, and utilize the poly learning rate policy [55] with the power of 5.0. The transformer stream parameters are initialized from Swin-S [47]. The model converges after 90k iterations with the batch size of 24. During testing, input images are resized to 256×256 before being fed into the network, and the outputs are resized to the original size.

4.2 Ablation study

To demonstrate the effectiveness of each key component in the proposed TGDNet (i.e., TGDE, SICAM, and PSAM), we conduct ablation experiments. The baseline model keeps the identical decoder structure with the full TGDNet, but removes the transformer stream, PSAM, and SICAM. For the fairness of comparison, the experimental settings keep constant except for the ablated parts.

The above model variants are quantitatively analyzed by MAE, F-measure, and S-measure on ORSSD and EORSSD, respectively, as shown in Table 1. The proposed TGDNet achieves the state-of-the-art performance on two datasets compared to other four variants. Compared to the baseline network (first row), by simply introducing the transformer (second row), the performance in terms of MAE, F-measure and S-measure reached 60.8%, 11.2% and 6.8% gain percentage on ORSSD, and 59.4%, 14.4% and 10.5% on EORSSD, respectively. It confirms the superiority of the extracted global information by transformer stream. Then, in order to evaluate the role of SICAM and PSAM separately, we insert SICAM and PSAM over the network structure in the second row, respectively. After adding SICAM, as shown in the third row, the percentage gain reaches 20.0%, 0.6%, and 1.4% in terms of MAE, F-measure, and S-measure on ORSSD. And it reaches 9.2%, 1.5%, and 1.4% on EORSSD, illustrating that SICAM improves the metrics in a balanced way. On the other hand, the network with embedded PSAM, as shown in the fourth row, the percentage gain reaches 32.0% and 1.7% in terms of MAE and S-measure on ORSSD, and reaches 27.6% and 1.9% on EORSSD. This

[1] https://li-chongyi.github.io/proj_optical_saliency.html.

[2] https://github.com/rmcong/EORSSD-dataset.

Table 1 Ablation experiments for the proposed TGDNet on ORSSD and EORSSD

Baseline	Transformer	SICAM	PSAM	ORSSD			EORSSD		
				$MAE \downarrow$	$maxF_\beta \uparrow$	$S_m \uparrow$	$MAE \downarrow$	$maxF_\beta \uparrow$	$S_m \uparrow$
✓				0.0255	0.8296	0.8644	0.0187	0.7622	0.8198
✓	✓			0.0100	0.9228	0.9236	0.0076	0.8698	0.9058
✓	✓	✓		0.0080	0.9282	0.9361	0.0069	0.8830	0.9189
✓	✓		✓	0.0068	0.9220	0.9395	0.0055	0.8816	0.9229
✓	✓	✓	✓	**0.0065**	**0.9365**	**0.9437**	**0.0049**	**0.8964**	**0.9286**

Bold indicates the top scores

verifies the effectiveness of PSAM in enhancing detail accuracy. The last row is the TGDNet with complete structure, which has the best performance among three metrics in all variants. It outperforms the baseline network in terms of MAE, F-measure, S-measure by 74.5%, 12.9%, 9.2% on ORSSD, and 73.8%, 17.6%, 13.3% on EORSSD, respectively.

The visual comparison is shown in Fig. 7. As can be seen, compared to the baseline model shown in Fig. 7c, the model with the transformer stream shown in Fig. 7d captures the more accurate object location. It reveals the advantage of transformer in capturing global information. as shown in Fig. 7e, by introducing SICAM, semantic information is further extracted, salient objects get more accurately located. In addition, the edge information is enhanced after embedding PSAM, as illustrated in Fig. 7f. Finally, as shown in Fig. 7g, the full TGDNet provides complete regions and detail structure for multi-scale objects. Therefore, PSAM and SICAM designed separately for the low-level and high-level features could allow our method to possess the robustness to accurately discover salient objects, sharply segment the object edges, and effectively suppress background noise. At this point, the effectiveness of each key component is demonstrated by both quantitative and qualitative results.

4.3 Comparison with state-of-the-art methods

We compare our method with 17 state-of-the-art SOD models on the testing subset of ORSSD and EORSSD in terms of qualitative and quantitative comparisons. These compared models include five traditional methods for NSIs (i.e., SO [56], RCRR [57], MR [58], GS [59], and SF [60]), six CNN-based methods for NSIs (i.e., DSS [25], EGNet [20], PoolNet [21], SCRN [61], BASNet [28], and MINet [18]), two transformer-based methods for NSIs (i.e., VST [48] and TransformerSOD [14]), and four CNN-based methods for optical RSIs (i.e., LVNet [34], MJRBM [37], DAFNet [12], and EMFINet [36]). Considering the fairness of the comparison, the experiment results are generated directly from the source code released by the authors, and the deep learning-based methods are retrained using the same training dataset (i.e., EORSSD) as ours.

4.3.1 Qualitative comparison

Some visual results of different methods are illustrated in Fig. 8. From the visual comparison, our method has significant advantages in multiple aspects. Firstly, our method performs better than traditional and CNN-based methods in terms of completeness and accuracy of locating salient objects (e.g., the $1 \sim 3$-rd rows), where our model achieves clear and complete detection in scenarios such as detecting rivers and playgrounds. Secondly, in detecting multiple and small objects (e.g., the $4 \sim 6$-th rows), our method outperforms traditional and CNN-based methods as well as transformer-based methods, where our model could recognize the quantity of salient objects, provides complete shape and clear edges. Finally, in suppressing cluttered background interference(e.g., the 7-th and 8-th rows), our method is more effective than other methods, where our model successfully suppresses the obstruction of background and detects salient objects without false detection of the background.

4.3.2 Quantitative comparison

To further evaluate the performance of these methods in a fair and comprehensive manner, all results are measured by: PR curves, F-measure curves, MAE, F-measure, and S-measure. We show the PR curves and F-measure curves of different methods on ORSSD and EORSSD datasets in Fig. 9. Our TGDNet is at the outermost (solid red line) in Fig. 9a–d, demonstrating the superior performance in both datasets.

For visual representation, we report the metrics in terms of MAE, F-measure, and S-measure in Table 2. Our method achieves best performance ranked the first at least in terms of two metrics on ORSSD and EORSSD, which is consistent with the results shown in Fig. 9. It is evident that the traditional methods lag far behind the deep learning-based methods on all three metrics. Methods designed

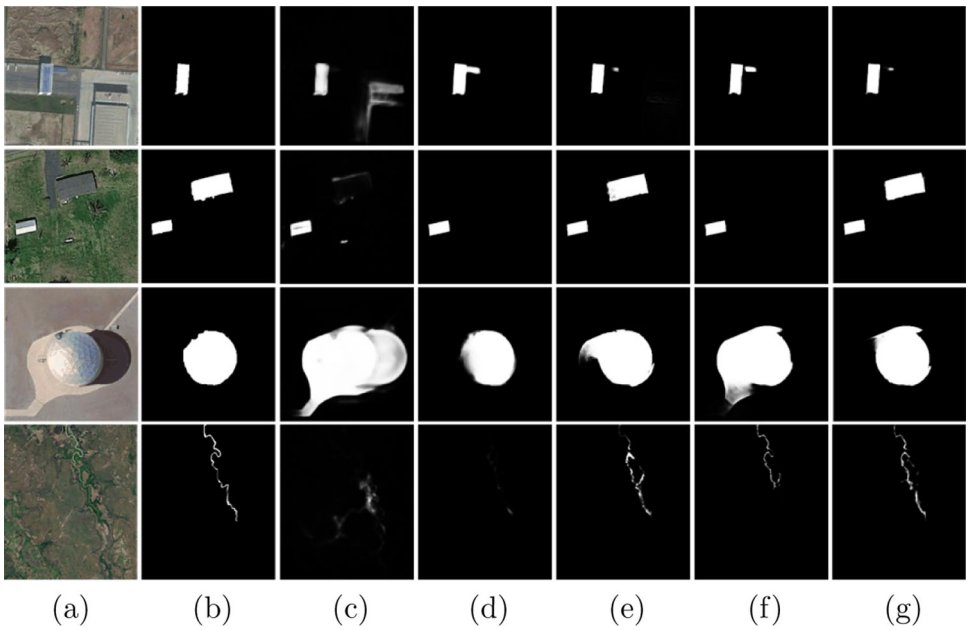

Fig. 7 Qualitative visual comparison of ablation study. **a** Optical RSIs. **b** GT. **c** Baseline. **d** Baseline+Transformer. **e** Baseline+Transformer+SICAM. **f** Baseline+Transformer+PSAM. **g** Baseline+Transformer+SICAM+PSAM (the full TGDNet)

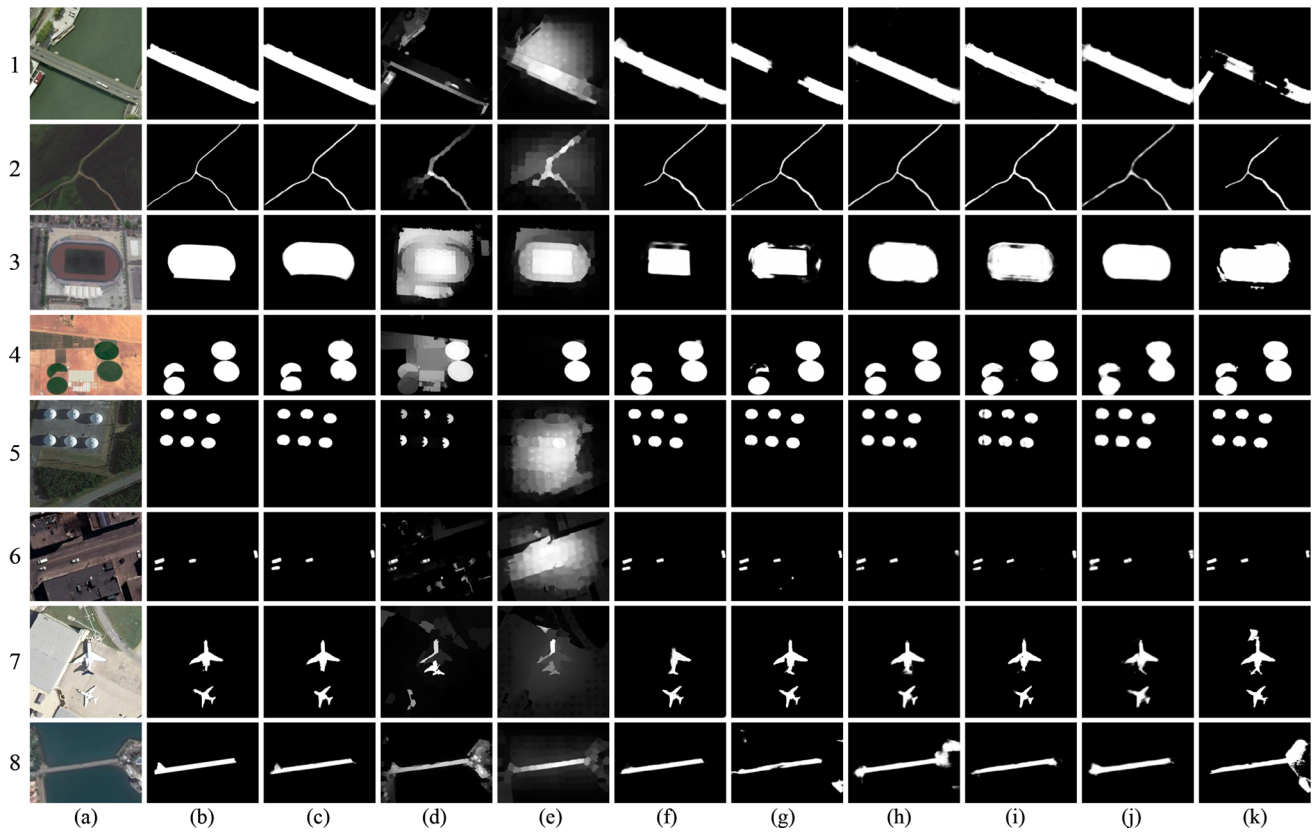

Fig. 8 Visual comparison of the proposed method with the other eight methods. **a** Optical RSIs. **b** GT. **c** Ours. **d** SO [56]. **e** RCRR [57]. **f** BASNet [28]. **g** MINet [18]. **h** VST [48]. **i** TransformerSOD [14]. **j** DAFNet [12]. **k** EMFINet [36]

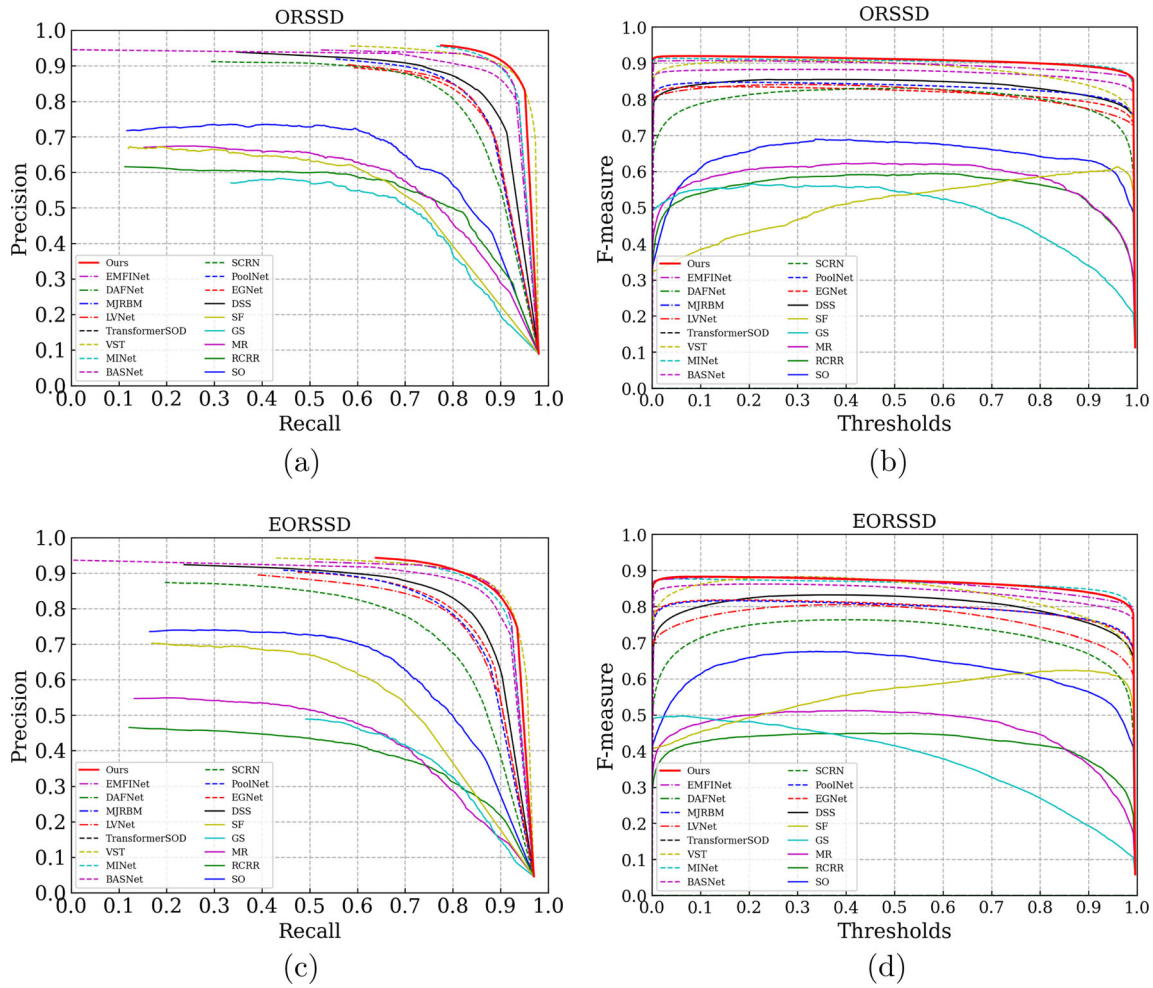

Fig. 9 Illustration of PR curves and F-measure curves on the testing subset of ORSSD and EORSSD of different methods. **a** PR curves on ORSSD. **b** F-measure curves on ORSSD. **c** PR curves on EORSSD. **d** F-measure curves on EORSSD. Better viewed in color

specifically for optical RSIs (LVNet [34], MJRBM [37], DAFNet [12], and EMFINet [36]) generally outperform CNN-based methods for NSIs retrained by optical RSIs, which illustrates the necessity of designing specific methods to address the unique properties of optical RSIs. Moreover, our method is better than the transformer-based methods (VST [48] and TransformerSOD [14]), especially in terms of MAE.

This illustrates that our method is brilliant in predicting the value of the single pixel. Compared with the second best method, in terms of MAE, our method achieves a percentage gain of 18.8% on ORSSD and 18.3% on EORSSD. In terms of F-measure, our method reaches a percentage gain of 1.8% on ORSSD and 0.5% on EORSSD. And in terms of S-measure, our method also achieves competitive performance. All these metrics demonstrate the effectiveness of our method.

4.4 Limitation

Our proposed SOD method achieves high accuracy. However, its dual-branch structure results in high computational complexity and time costs, as shown in Table 3. Compared with state-of-the-art methods, our method requires a higher running time and FLOPs, which could limit its practicality for real-time applications and resource-constrained devices. Future work will focus on developing lightweight models to stride a balance between performance and practicality.

5 Conclusion

This paper presents a transformer guidance dual-stream network (TGDNet) for SOD in optical RSIs. Guided by the long-range dependence property of transformer, the proposed TGDNet could capture fantastic encoding features

Table 2 Quantitative comparison of our proposed TGDNet with other 16 methods on testing subset of ORSSD and EORSSD

Method	ORSSD			EORSSD		
	$MAE \downarrow$	$maxF_\beta \uparrow$	$S_m \uparrow$	$MAE \downarrow$	$maxF_\beta \uparrow$	$S_m \uparrow$
SO [56]	0.0626	0.6598	0.7660	0.0465	0.6337	0.7417
RCRR [57]	0.1277	0.5591	0.6850	0.1647	0.4018	0.6012
MR [58]	0.1355	0.5657	0.6722	0.1591	0.4499	0.5956
GS [59]	0.1248	0.5254	0.6577	0.1365	0.4346	0.5827
SF [60]	0.0760	0.5816	0.6418	0.0400	0.5763	0.6774
DSS [25]	0.0264	0.8527	0.8734	0.0160	0.8361	0.8767
EGNet [20]	0.0267	0.8274	0.8713	0.0150	0.8231	0.8724
PoolNet [21]	0.0287	0.8319	0.8716	0.0167	0.8143	0.8671
SCRN [61]	0.0401	0.7879	0.8346	0.0247	0.7168	0.8089
BASNet [28]	0.0172	0.8741	0.9175	0.0112	0.8512	0.9157
MINet [18]	0.0128	0.9119	0.9213	0.0078	0.8830	0.9193
VST [48]	***0.0104***	0.9072	0.9331	***0.0066***	***0.8866***	***0.9211***
TransformerSOD [14]	**0.0080**	**0.9197**	***0.9335***	0.0073	0.8805	0.9176
LVNet [34]	0.0207	0.8263	0.8813	0.0145	0.7824	0.8650
MJRBM [37]	0.0146	0.8885	0.9199	0.0099	0.8555	0.9091
DAFNet [12]	0.0125	***0.9174***	0.9191	**0.0060**	**0.8922**	0.9167
EMFINet [36]	0.0109	0.9002	**0.9366**	0.0084	0.8720	*0.9290*
TGDNet	*0.0065*	*0.9365*	*0.9437*	*0.0049*	*0.8964*	**0.9286**

Italic, **Bold**, and ***Bolditalic*** indicate the top three scores, respectively

Table 3 Running time and FLOPs of state-of-the-art methods (SCRN [61], PoolNet [21], DSS [25], MINet [18], BASNet [28], EGNet [20], VST [48], and TransformerSOD [14]) and the proposed TGDNet

Methods	SCRN	PoolNet	DSS	MINet	BASNet	EGNet	VST	TransformerSOD	TGDNet
Time (m)	0.013	0.006	0.009	0.009	0.008	0.007	0.010	0.011	0.028
FLOPs (G)	6.10	38.14	48.71	71.67	97.54	120.80	23.16	47.51	51.38

that aggregates global semantic features and local detail information. In addition, we investigate the sequence inheritance channel attention module (SICAM), which allows the obtained attention weights to match multiple feature maps simultaneously, thereby extracting and refining the multi-scale information of high-level features. Meanwhile, we design a pyramid spatial attention module (PSAM) to enhance the detail information of low-level features. Extensive experiments on two benchmark datasets validate that the proposed model outperforms other state-of-the-art methods under different evaluation metrics.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant 62171315.

Data Availability The datasets and materials used or analyzed during the current study are available from the corresponding author on reasonable request.

Code Availability The code used or analyzed during the current study is available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20(11):1254–1259
2. Borji A, Cheng M-M, Hou Q, Jiang H, Li J (2019) Salient object detection: a survey. Comput Visual Media 5(2):117–150
3. Mohamed IS, Capitanelli A, Mastrogiovanni F, Rovetta S, Zaccaria R (2020) Detection, localisation and tracking of pallets using machine learning techniques and 2D range data. Neural Comput Appl 32(13):8811–8828
4. Wang X, You S, Li X, Ma H (2018) Weakly-supervised semantic segmentation by iteratively mining common object features. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1354–1362

5. Wang W, Shen J, Porikli F, Yang R (2019) Semi-supervised video object segmentation with super-trajectories. IEEE Trans Pattern Anal Mach Intell 41(04):985–998
6. Das A, Agrawal H, Zitnick L, Parikh D, Batra D (2017) Human attention in visual question answering: do humans and deep networks look at the same regions? Comput Vis Image Underst 163:90–100
7. Fang H, Gupta S, Iandola F, Srivastava RK, Deng L, Dollár P, Gao J, He X, Mitchell M, Platt JC et al (2015) From captions to visual concepts and back. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1473–1482
8. Mechrez R, Shechtman E, Zelnik-Manor L (2019) Saliency driven image manipulation. Mach Vis Appl 30(2):189–202
9. Li W, Zhu H, Yang S, Wang P, Zhang H (2022) GA-SRN: graph attention based text-image semantic reasoning network for fine-grained image classification and retrieval. Neural Comput Appl. https://doi.org/10.1007/s00521-022-07617-3
10. Ma X, Zhao R, Shi Z (2020) Multiscale methods for optical remote-sensing image captioning. IEEE Geosci Remote Sens Lett. https://doi.org/10.1109/LGRS.2020.3009243
11. Han Y, Yang X, Pu T, Peng Z (2021) Fine-grained recognition for oriented ship against complex scenes in optical remote sensing images. IEEE Trans Geosci Remote Sens. https://doi.org/10.1109/TGRS.2021.3123666
12. Zhang Q, Cong R, Li C, Cheng M-M, Fang Y, Cao X, Zhao Y, Kwong S (2020) Dense attention fluid network for salient object detection in optical remote sensing images. IEEE Trans Image Process 30:1305–1317
13. Yang Q, Zhou Y, Chai X, Zhang M, Zhang W, Wang J (2022) Exploring class-agnostic pixels for scribble-supervised high-resolution salient object detection. Neural Comput Appl. https://doi.org/10.1007/s00521-022-07915-w
14. Mao Y, Zhang J, Wan Z, Dai Y, Li A, Lv Y, Tian X, Fan D-P, Barnes N (2021) Transformer transforms salient object detection and camouflaged object detection. arXiv preprint arXiv:2104.10127
15. Liu Z, Wang Y, Tu Z, Xiao Y, Tang B (2021) TriTransNet: RGB-D salient object detection with a triplet transformer embedding network. In: Proceedings of the 29th ACM international conference on multimedia, pp 4481–4490
16. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
17. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M (2021) Transformers in vision: a survey. ACM Comput Surv (CSUR)
18. Pang Y, Zhao X, Zhang L, Lu H (2020) Multi-scale interactive network for salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9413–9422
19. Chen T, Hu X, Xiao J, Zhang G, Wang S (2022) CFIDNet: cascaded feature interaction decoder for RGB-D salient object detection. Neural Comput Appl 34(10):7547–7563
20. Zhao J-X, Liu J-J, Fan D-P, Cao Y, Yang J, Cheng M-M (2019) EGNet: edge guidance network for salient object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8779–8788
21. Liu J-J, Hou Q, Cheng M-M, Feng J, Jiang J (2019) A simple pooling-based design for real-time salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3917–3926
22. Chen T, Xiao J, Hu X, Zhang G, Wang S (2022) Spatiotemporal context-aware network for video salient object detection. Neural Comput Appl. https://doi.org/10.1007/s00521-022-07330-1
23. Su J, Li J, Zhang Y, Xia C, Tian Y (2019) Selectivity or invariance: boundary-aware salient object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3799–3808
24. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440
25. Hou Q, Cheng M-M, Hu X, Borji A, Tu Z, Torr PH (2017) Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3203–3212
26. Xie S, Tu Z (2015) Holistically-nested edge detection. In: Proceedings of the IEEE international conference on computer vision, pp 1395–1403
27. Liu N, Han J, Yang M-H (2018) Picanet: learning pixel-wise contextual attention for saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3089–3098
28. Qin X, Zhang Z, Huang C, Gao C, Dehghan M, Jagersand M (2019) Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7479–7489
29. Zhao T, Wu X (2019) Pyramid feature attention network for saliency detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3085–3094
30. Siris A, Jiao J, Tam GK, Xie X, Lau RW (2021) Scene context-aware salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4156–4166
31. Wu Z, Su L, Huang Q (2021) Decomposition and completion network for salient object detection. IEEE Trans Image Process 30:6226–6239
32. Zhao D, Wang J, Shi J, Jiang Z (2015) Sparsity-guided saliency detection for remote sensing images. J Appl Remote Sens 9(1):095055
33. Zhang L, Liu Y, Zhang J (2019) Saliency detection based on self-adaptive multiple feature fusion for remote sensing images. Int J Remote Sens 40(22):8270–8297
34. Li C, Cong R, Hou J, Zhang S, Qian Y, Kwong S (2019) Nested network with two-stream pyramid for salient object detection in optical remote sensing images. IEEE Trans Geosci Remote Sens 57(11):9156–9166
35. Li C, Cong R, Guo C, Li H, Zhang C, Zheng F, Zhao Y (2020) A parallel down-up fusion network for salient object detection in optical remote sensing images. Neurocomputing 415:411–420
36. Zhou X, Shen K, Liu Z, Gong C, Zhang J, Yan C (2021) Edge-aware multiscale feature integration network for salient object detection in optical remote sensing images. IEEE Trans Geosci Remote Sens
37. Tu Z, Wang C, Li C, Fan M, Zhao H, Luo B (2021) ORSI salient object detection via multiscale joint region and boundary model. IEEE Trans Geosci Remote Sens. https://doi.org/10.1109/TGRS.2021.3101359
38. Li G, Liu Z, Lin W, Ling H (2021) Multi-content complementation network for salient object detection in optical remote sensing images. IEEE Trans Geosci Remote Sens. https://doi.org/10.1109/TGRS.2021.3131221
39. Cong R, Zhang Y, Fang L, Li J, Zhang C, Zhao Y, Kwong S (2021) RRNet: relational reasoning network with parallel multi-scale attention for salient object detection in optical remote sensing images. IEEE Trans Geosci Remote Sens. https://doi.org/10.1109/TGRS.2021.3123984
40. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European conference on computer vision, pp 213–229. Springer, Berlin

41. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
42. Ye L, Rochan M, Liu Z, Wang Y (2019) Cross-modal self-attention network for referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10502–10511
43. Ranftl R, Bochkovskiy A, Koltun V (2021) Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 12179–12188
44. Zheng S, Lu J, Zhao H, Zhu X, Luo Z, Wang Y, Fu Y, Feng J, Xiang T, Torr PH et al (2021) Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6881–6890
45. Wang H, Zhu Y, Adam H, Yuille A, Chen L-C (2021) Max-deeplab: end-to-end panoptic segmentation with mask transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5463–5474
46. Chen H, Wang Y, Guo T, Xu C, Deng Y, Liu Z, Ma S, Xu C, Xu C, Gao W (2021) Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12299–12310
47. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030
48. Liu N, Zhang N, Wan K, Shao L, Han J (2021) Visual saliency transformer. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4722–4732
49. Xu C, Liu X, Zhao W (2022) Attention-guided salient object detection using autoencoder regularization. Appl Intell. https://doi.org/10.1007/s10489-022-03917-2
50. Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19
51. Zhang L, Ma J (2021) Salient object detection based on progressively supervised learning for remote sensing images. IEEE Trans Geosci Remote Sens 59(11):9682–9696
52. Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency-tuned salient region detection. In: 2009 IEEE conference on computer vision and pattern recognition, pp 1597–1604. IEEE
53. Fan D-P, Cheng M-M, Liu Y, Li T, Borji A (2017) Structure-measure: a new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision, pp 4548–4557
54. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
55. You Y, Gitman I, Ginsburg B (2017) Scaling sgd batch size to 32k for imagenet training. arXiv preprint arXiv:1708.03888
56. Zhu W, Liang S, Wei Y, Sun J (2014) Saliency optimization from robust background detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2814–2821
57. Yuan Y, Li C, Kim J, Cai W, Feng DD (2017) Reversion correction and regularized random walk ranking for saliency detection. IEEE Trans Image Process 27(3):1311–1322
58. Yang C, Zhang L, Lu H, Ruan X, Yang M-H (2013) Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3166–3173
59. Wei Y, Wen F, Zhu W, Sun J (2012) Geodesic saliency using background priors. In: European conference on computer vision, pp 29–42. Springer, Berlin
60. Perazzi F, Krähenbühl P, Pritch Y, Hornung A (2012) Saliency filters: contrast based filtering for salient region detection. In: 2012 IEEE conference on computer vision and pattern recognition, pp 733–740. IEEE
61. Wu Z, Su L, Huang Q (2019) Stacked cross refinement network for edge-aware salient object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 7264–7273