Hybrid Feature Aligned Network for Salient Object Detection in Optical Remote Sensing Imagery

Qi Wang, *Senior Member, IEEE,* Yanfeng Liu, *Student Member, IEEE,* Zhitong Xiong, *Member, IEEE,* and Yuan Yuan, *Senior Member, IEEE*

***Abstract*—Recently, salient object detection in optical remote sensing images (RSI-SOD) has attracted great attention. Benefiting from the success of deep learning and the inspiration of natural SOD task, RSI-SOD has achieved fast progress over the past two years. However, existing methods usually suffer from the intrinsic problems of optical RSIs, 1) cluttered background; 2) scale variation of salient objects; 3) complicated edges and irregular topology. To remedy these problems, we propose a hybrid feature aligned network (HFANet) jointly modeling boundary learning to detect salient objects effectively. Specifically, we design a hybrid encoder by unifying two components to capture global context for mitigating the disturbance of complex background. Then, to detect multiscale salient objects effectively, we propose a Gated Fold-ASPP (GF-ASPP) to extract abundant context in the deep semantic features. Furthermore, an adjacent feature aligned module (AFAM) is presented for integrating adjacent features with unparameterized alignment strategy. Finally, we propose a novel interactive guidance loss (IGLoss) to combine saliency and edge detection, which can adaptively perform mutual supervision of the two sub-tasks to facilitate detection of salient objects with blurred edges and irregular topology. Adequate experimental results on three optical RSI-SOD datasets reveal that the presented approach exceeds 18 state-of-the-art ones. All codes and detection results are available at *https://github.com/lyf0801/HFANet*.**

***Index Terms*—Salient object detection, optical remote sensing image, feature alignment, multiscale context modeling, jointing boundary learning.**

I. Introduction

SALIENT object detection (SOD) is the task of locating the most visually attractive objects/regions and generating pixel-wise saliency maps in images/videos [1]–[4]. SOD for natural scene images (NSIs) has obtained great success benefiting from fully convolutional neural networks (FCN) in recent years. It can be applied in numerous research areas successfully, such as image/video segmentation [5]–[7], traffic sign detection [8]–[10], and object tracking [11].

Inspired by SOD task in NSIs, some researchers propose the task of SOD in optical remote sensing images (RSIs) [12]–[14], namely, RSI-SOD. It aims at extracting airplanes, vehicles, storage tanks, islands, buildings, rivers and other objects (e.g., Fig. 1) which attract human attention in optical RSIs. Note that optical RSIs only with RGB bands are dissimilar to hyperspectral RSIs that contain more spectral information [15]. Besides, SOD is distinguished from common object detection (OD) [16], [17] and anomaly detection (AD) [18]. Firstly, OD is a general task to discover all objects, while AD merely identifies abnormal objects (anomaly do not necessarily cause saliency). Secondly, the latter two always utilize bounding boxes for supervision/prediction, while SOD performs pixel-wise learning. As illustrated in [12]–[14], RSI-SOD offers a pre-processing technique to promote other downstream vision tasks, such as scene classification [19], image captioning [20] and image fusion [21].

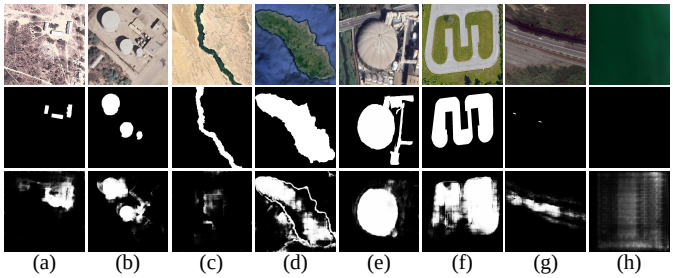


Fig. 1. Visual examples of typical optical RSIs (top row), GT (middle row), and corresponding saliency maps by NLDF method (bottom row).

In recent years, convolutional neural networks (CNNs) have significantly stimulated NSI-SOD, and numerous excellent algorithms have been proposed [22]–[31]. Unlike NSIs taken by the photographers with handheld cameras [13], optical RSIs are gathered from high-altitude top-down angle by satellites or other aerial platforms, which results in wide imaging view and cluttered background. Furthermore, due to the variety, size, and arrangement diversity of man-made salient objects, it also leads to some challenges 1) cluttered background; 2) extreme scale variation; 3) complicated edges and irregular topology of salient objects in optical RSIs [14]. There are some visual examples of optical RSIs and the corresponding ground-truths are illustrated in Fig. 1. As presented, it is difficult to extract foreground salient objects in Figs. 1(a) and 1(b) because of complex background. The salient objects in Figs. 1(c)–1(f), i.e., river, island, and buildings, have severe irregular topology and complicated boundaries, which further hinder the detection performance of RSI-SOD. Besides, some optical RSIs have extremely small objects or even no salient regions, as shown in

Manuscript received April 7, 2022; revised May 12, 2022; accepted June 2, 2022. This work was supported by the National Natural Science Foundation of China under Grant U21B2041, Grant U1864204, and Grant 61825603. *(Corresponding author: Yuan Yuan.)*

Qi Wang and Yuan Yuan are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: crabwq@gmail.com, y.yuan1.ieee@gmail.com).

Yanfeng Liu is with the School of Computer Science and School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: liuyanfeng99@gmail.com).

Zhitong Xiong is with the Data Science in Earth Observation (SiPEO, former: Signal Processing in Earth Observation), Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: zhitong.xiong@tum.de).

Figs. 1(g) and 1(h). These diverse scene patterns and particular characteristics of optical RSIs make NSI-SOD methods unable to obtain satisfactory results by a direct transplantation (shown in bottom row of Fig. 1). Therefore, it is highly necessary to design specialized SOD approaches for optical RSIs.

To mitigate the above-mentioned challenges, previous work has made progress inspired by NSI-SOD. For instance, three separate datasets and benchmarks of some existing NSI-SOD algorithms are proposed in [12]–[14]. Most of them [32]–[36] design various multiscale feature fusion strategies of encoder/decoder to improve the ability for modeling the scale variation of salient objects, and EMFINet [36] even adopts image pyramid strategy. SARNet [35], MJRBM [14] and EMFINet [36] all perform joint learning frameworks for salient regions and boundaries, which improve the detection performance to a certain extent. In addition, there are several approaches [14], [34], [35] proposing visual attention mechanisms to interpret optical RSIs with cluttered background. The above-mentioned studies have promoted the development of RSI-SOD, but still remain the following problems:

1) Lack of capability to extract global context information from optical RSIs. The previous models are all based on CNNs, which only have local receptive fields and cannot restrain the global disturbance of background clutters effectively.

2) Multiscale feature misalignment and inadequate context modeling problems. Especially, existing studies only design some feature fusion strategies without considering the spatial inconsistency issue between different features.

3) Suffering from complicated or blurred boundaries of salient objects. Existing methods [14], [35], [36] cannot make better use of the edge cues to detect salient regions effectively.

Thus more research efforts are still required to handle these specialized issues for RSI-SOD. To this end, we propose a hybrid feature aligned network (HFANet) combining edge learning to detect salient regions accurately in optical RSIs. To enable our model to capture the global context, we design a hybrid encoder, which enhances the representation learning for foreground objects in RSIs with background interference. The hybrid encoder combines the advantages of CNNs and Transformer architectures in an eclectic manner. It utilizes the inductive bias introduced in CNNs for effective local context modeling in the shallow layers, and introduces Transformer blocks in the deep layers for a long-distance context learning. Such a structure enables our model to adaptively aggregate global and local context information at different layers. To our best knowledge, it is the first time that Transformer blocks are introduced to guide deep feature extraction in RSI-SOD task. Since the misalignment problem between multiscale features is always neglected in previous studies, we present an adjacent feature aligned module (AFAM) to alleviate this inconsistency. AFAM takes two adjacent feature maps with different scales as input, and uses upsampling and element-wise operations as pre-processing to obtain preliminary aggregated features. Then it adopts a learnable alignment strategy to generate final coalescent features. In addition, we present a gated fold atrous spatial pyramid pooling (GF-ASPP) to precisely localize salient regions of various scales in RSIs inspired by [29]. Finally, to facilitate saliency detection by better exploring edge cues, we propose an interactive guidance cross-entropy loss function (IGLoss), which can dynamically adjust the weights of various pixels during training process with mutually supervised salient object and edge detection. The contributions of this article are summarized as follows.

- To handle the aliasing problem caused by cluttered background of optical RSIs, we propose a hybrid encoder that combines CNNs and Transformer architectures to learn local/global context from different levels adaptively.
- To detect multiscale salient objects effectively, we design AFAM to address the misalignment issue between adjacent features. Additionally, GF-ASPP is introduced to capture abundant deep contextual features.
- We present IGLoss to recognize objects with complex edges and irregular topology in optical RSIs. It adopts an adaptive approach to joint region and boundary detection in a mutual supervised manner.
- Towards a fair comparison of existing methods, we first release a public code library to foster future research. It includes 2 traditional methods, 10 CNN-based algorithms for NSI-SOD, and 7 CNN-based methods for RSI-SOD.
- Extensive experiments show that our algorithm outperforms 18 state-of-the-art approaches, and we find that the proposed IGLoss is more applicable to optical RSIs than recent novel objective functions [37]–[39].

II. Related Work

We separately summarize the traditional and deep learning-based work of SOD for NSIs and optical RSIs in this section.

A. SOD in Natural RGB Images

The first visual saliency system is achieved by Itti *et al.* in the seminal work [1]. After that, early approaches based on low-level features are proposed, which are inefficient and lack semantic information. With the great success of deep learning, significant efforts have been made by FCN-based models.

1) Traditional Approaches: From the initial biological inspiration, many algorithms based on mathematics, statistics or information theory have been proposed to make better predictions for human attention [1]–[4], [40]–[42]. These methods always extract low-level features such as color, texture, brightness, scale, orientation, boundary or even relative position of objects from images. Among them, Achanta *et al.* propose a simple model by using features of color and luminance to estimate center-surround contrast [3]. Zhu *et al.* present a novel differential threshold-based psychologic feature and unify it into a supervised Markov Random Field framework [41]. Wang *et al.* explicitly consider SOD as a multiple instance learning problem and adopt low-, mid-, high-level features for training and testing [4]. However, the models only leverage handcrafted visual features, and always fail in cluttered patterns since they suffer from a deficiency of deep contextual knowledge.

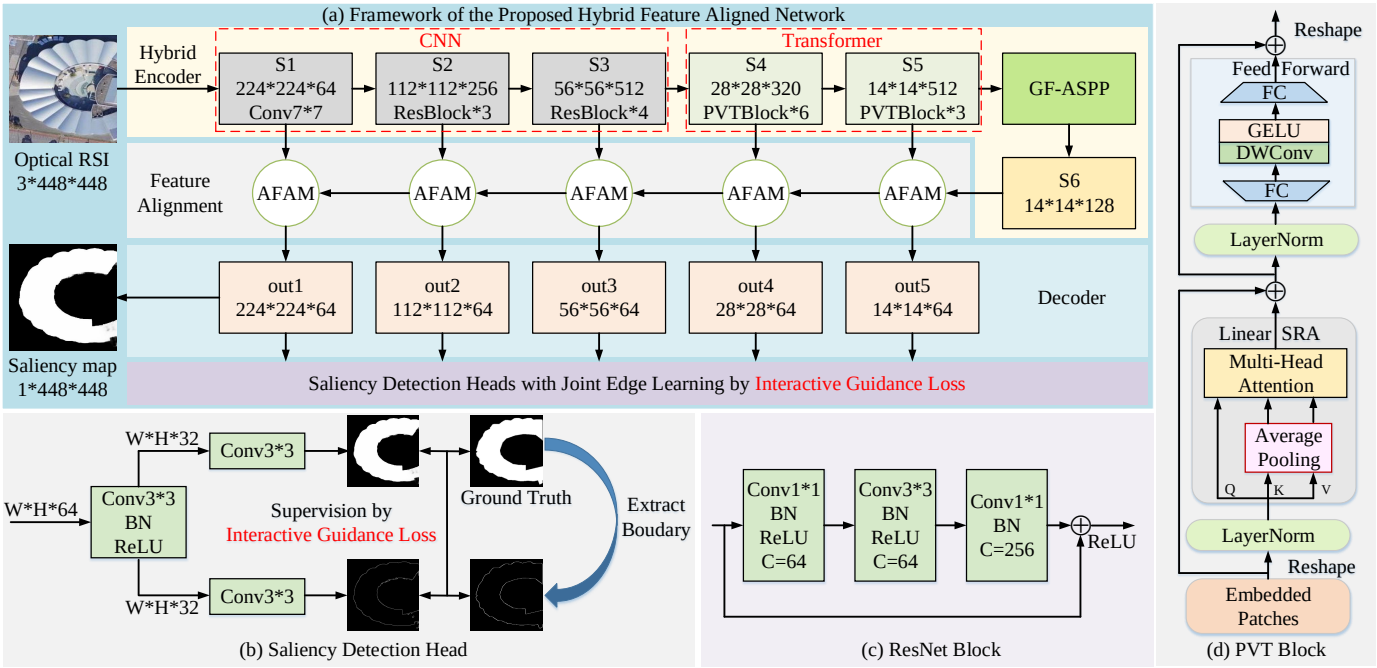


Fig. 2. (a) Framework of HFANet. (b) Framework of the saliency detection head. (c) Framework of Residual Block. (d) Framework of PVT block.

2) Deep Learning-based Approaches: In the last six years, various CNN-based SOD approaches have been extensively explored [22]–[31]. These approaches could be roughly classified into three types, i.e., patch-wise, pixel-wise and visual attention models. A typical patch-wise method [43] utilizes a two-stage network for local estimation and global search, respectively, and generates final saliency map by summing of salient regions. Although these models outperform traditional methods, the patch-based strategies still cannot extract sufficient spatial information from NSIs. Pixel-wise SOD algorithms adopt FCN-based networks to predict saliency for each pixel of an image, and the mainstream deep learning-based methods are all pixel-wise models [24]–[26]. Typically, Pang *et al.* [27] propose MINet to characterize the multiscale information by building two multi-path feature interaction models. F3Net [30] has realized that skip feature fusion of CNNs may bring negative effects and designs related modules to reduce the impact of this problem as much as possible. PFSNet [31] shrinks adjacent features hierarchically to avoid leaping feature fusion operations and reduce noise. Recently, some researchers have applied visual attention mechanisms to predict saliency, such as [24], [26], [44]. For instance, Chen *et al.* apply reverse attention into residual learning to guide SOD [24]. Compared with conventional evaluation strategies [3], [45], a novel metric named *S-measure* [46] is designed to gauge the structure information for saliency detection. Furthermore, some effective objective loss functions are presented to train the networks stably and accurately [37]–[39]. These models and functions greatly facilitate the development of NSI-SOD. Although there are many disadvantages in applying them directly to optical RSIs with complex scenes, they still provide a number of baselines and inspirations for RSI-SOD.

B. SOD in Optical RSIs

RSI-SOD aims to explore the most attractive objects/regions from optical RSIs. Benefiting from the success of NSI-SOD, RSI-SOD has made some progress in recent years. We provide a briefly review about the traditional and CNN-based methods for RSI-SOD in this subsection.

1) Traditional Approaches: As pioneering work, Zhao *et al.* propose two private datasets, and utilize global- and context-based dictionaries to generate sparse representations for RSIs [47]. After that, Zhang *et al.* propose several unsupervised algorithms [48]–[50], such as applying super-pixel, statistical saliency feature or color information content to obtain saliency maps of residential areas, etc. To explore the intrinsic relationship between various cues, an adaptive multi-feature fusion saliency method is proposed for region of interest extraction in optical RSIs [50]. Huang *et al.* present a contrast-weighted term to constrain contrast-weighted patterns from appearance in learned saliency dictionaries [51]. These methods have gained less research attention in recent years because of the rise of deep learning.

2) Deep Learning-based Approaches: Li *et al.* propose the first public dataset, namely ORSSD, and evaluate it using a well-designed CNN-based model [12]. Based on this work, RSI-SOD has attracted more research attention in recent years. To promote the benchmarks of this field, Zhang *et al.* and Tu *et al.* propose two more impressive public datasets named EORSSD [13] and ORSI-4199 [14], respectively. To handle the imbalanced scale variation of salient objects, existing algorithms always present multiscale feature integration models [32]–[35]. It is worth noting that EMFINet [36] integrates three strategies of image pyramid, feature pyramid, and edge learning to enhance the detection performance. RRNet [34] first explores graph convolution networks for RSI-SOD,

and propose two relational reasoning models for space and channel separately. MCCNet [33] embeds multiple content complementary blocks to investigate the complementation of various features for RSI-SOD. However, there are still many deficiencies in the existing work. For example, these methods have only been evaluated on two datasets in their papers except MJRBM [14]. Besides, they still have difficulty in recognizing regions with complex edges and irregular topology. Furthermore, the feature misalignment issue and lack of global view of CNNs have not been considered in existing methods.

III. Methodology

In this section, we detail the presented HFANet with the IGLoss as shown in Fig. 2. The model adopts the classical encoder-decoder network paradigm, and the hybrid encoder is exploited for local and global feature extraction. To explore more sufficient multiscale context, we connect the proposed GF-ASPP at the end of the hybrid encoder. Additionally, we apply five AFAM blocks as the decoder to alleviate the spatial misalignment issue of the features at adjacent scales.

A. Hybrid Encoder for Local and Global Context Modeling

CNN-based models converge fast due to inductive bias but lack the ability to capture long-range dependencies, especially for optical RSIs. Transformers extract local/global contextual information adaptively with more flexible ability for modeling unstructured data [52]–[54]. However, they have a large amount of computation with slow convergence speed, and cannot perform well on small-scale datasets. Inspired by the studies that exploit the complementary properties of CNNs and Transformers [55]–[58], we seek a hybrid encoder capable of extracting rich local context in shallow stages as well as global context in deep stages for RSI-SOD.

CNNs have excellent local receptive fields in shallow layers, but lack long-range dependency modeling capacity in deep layers. To this end, we utilize the shallow blocks of ResNet50 to extract local features for optical RSIs. Formally, let the input RSI as $I \in \mathbb{R}^{3\times448\times448}$. We first apply a 7×7 convolution to extract shallow feature $S_1 \in \mathbb{R}^{64\times224\times224}$. As shown in Fig. 2(a), $S_2 \in \mathbb{R}^{256\times112\times112}$, $S_3 \in \mathbb{R}^{512\times56\times56}$ are the output features by stacking the residual blocks referring to stage 1 and 2 of ResNet50. Mathematically

$$\begin{aligned} S_2 &= \mathcal{F}^3_{\text{res}}(S_1), \\ S_3 &= \mathcal{F}^4_{\text{res}}(S_2), \\ e.g.,\ \mathcal{F}^1_{\text{res}}(X) &= \mathcal{K}_{1\times1}(\mathcal{K}_{3\times3}(\mathcal{K}_{1\times1}(X))) \oplus X, \end{aligned} \tag{1}$$

where $\mathcal{F}^i_{\text{res}}(\cdot)$ indicates the combined operations of i residual blocks, $\mathcal{K}_{1\times1}(\cdot)$ and $\mathcal{K}_{3\times3}(\cdot)$ denote 1×1 and 3×3 convolution with Batch-Normalization (BN) and ReLU layer.

Benefiting from the linear complexity of PVTv2 blocks [54], we employ them to equip stage 3 and 4 of the hybrid encoder at a small computational cost. PVTv2 block consists of a patch embedding layer, a linear spatial reduction attention (LSRA) layer and a convolutional feed-forward layer (CFFL). In the last two stages of hybrid encoder, PVTv2 blocks control the scale of the feature maps through a spatial shrinking strategy [54]. As illustrated in Fig. 2, the output features of stage 3 and 4 are $S_4 \in \mathbb{R}^{320\times28\times28}$, $S_5 \in \mathbb{R}^{512\times14\times14}$, i.e.,

$$\begin{aligned} S_4 &= \mathcal{F}^6_{\text{pvt}}(S_3), \\ S_5 &= \mathcal{F}^3_{\text{pvt}}(S_4), \\ e.g., \mathcal{F}^1_{\text{pvt}}(X) &= \mathcal{F}_{CFFL}(\mathcal{F}_{LRSA}(\mathcal{F}_{linear}(X)) \oplus X), \end{aligned} \tag{2}$$

where $\mathcal{F}^i_{\text{pvt}}(\cdot)$ denotes the combined operations of i PVTv2 blocks, $\mathcal{F}_{linear}(\cdot)$ indicates the function of linear projection, $\mathcal{F}_{LSRA}(\cdot)$ is the function of LSRA, and $\mathcal{F}_{CFFL}(\cdot)$ indicates the operation of CFFL. Suppose a query, a key, and a value as Q, K, $V \in \mathbb{R}^{(HW)\times C}$, respectively. We input them into LSRA, and the output is

$$\mathcal{F}_{LSRA}(Q,K,V) = \mathcal{F}_{cat}(h_0,...,h_N)W^A, \tag{3}$$

$$h_i = \mathcal{F}_{Atten}(QW^Q_i,\ \mathcal{F}_{Avg}(K)W^K_i,\ \mathcal{F}_{Avg}(V)W^V_i), \tag{4}$$

where H, W are the spatial height and width of the input before the attention operation, C is the dimension of patch embedding, and N is the head number of self-attention. W^A, W^Q_i, W^K_i, W^V_i are the linear projection parameters. $\mathcal{F}_{Avg}(\cdot)$ represents the operation of spatial average pooling. $\mathcal{F}_{cat}(\cdot)$ denotes the feature concatenation as in [59]. $\mathcal{F}_{Atten}(\cdot)$ is the self-attention function as follows:

$$\mathcal{F}_{Atten}(\mathbf{q},\mathbf{k},\mathbf{v}) = \text{softmax}(\frac{\mathbf{q}\mathbf{k}^{\mathrm{T}}}{\sqrt{d}})\mathbf{v}, \tag{5}$$

where $d = C/N$ is the dimension of each head. LSRA utilizes average pooling to shrink spatial scale, which can reduce the computational cost caused by self-attention. In this way, every element in $\mathbf{q}$ and $\mathbf{k}$ can obtain their pair-wise context production, and thus transformer block is able to extract long-distance dependencies with global receptive fields intrinsically. Furthermore, PVTv2 block integrates a 3×3 depth-wise convolution with GELU [60] activation layer into CFFL between two fully-connected layers, in which CFFL is illustrated as

$$\mathcal{F}_{CFFL}(X) = \mathcal{F}_{FC}(\mathcal{F}_{GELU}(\mathcal{F}_{DW}(\mathcal{F}_{FC}(X)))) \oplus X, \tag{6}$$

where $\mathcal{F}_{FC}(\cdot)$ indicates the operation of fully-connected layer, and $\mathcal{F}_{DW}(\cdot)$ denotes the 3×3 depth-wise convolution.

The hybrid encoder combines the strengths of convolution and self-attention strategies together in a simple manner. Different from UTNet [56], the proposed hybrid encoder applies two components in different stages separately. It ensures the integrity of the two components and increases the computational complexity as little as possible to enhance the modeling ability for long-range dependencies in deep layers.

B. Adjacent Feature Aligned Module (AFAM)

The feature misalignment issue in CNNs is rarely considered. As mentioned in [61], the downsampling operations and indiscriminate fusion strategies are the main reasons for this issue. We propose AFAM, which alleviates the spatial misalignment issue between adjacent features at various scales of RSIs. Unlike previous studies that employ a parameterized learning approach and align single feature [62]–[64], the presented AFAM progressively handles the inconsistency between

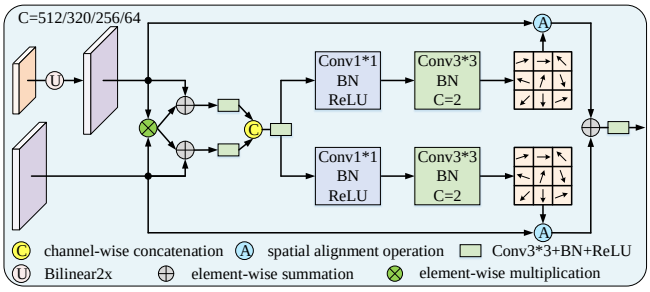


Fig. 3. The framework of the presented AFAM.

adjacent features and learn two offset maps in a non-parameter manner from the aggregated features.

Assume that $a \in \mathbb{R}^{C_1 \times H \times W}$, $b \in \mathbb{R}^{C_2 \times 2H \times 2W}$ indicate the features of adjacent stages of the encoder and decoder (e.g., S_5 and S_6, S_4 and out_5), AFAM upsamples a and uses element-wise multiplication to pay more attention to common elements between a and b. After that, we add the original features with the captured common elements to avoid inappropriate information redundancy, i.e.

$$\begin{aligned} a' &= \mathcal{K}_{3\times3}((\mathcal{F}_{up}(a) \otimes b) \oplus \mathcal{F}_{up}(a)), \\ b' &= \mathcal{K}_{3\times3}((\mathcal{F}_{up}(a) \otimes b) \oplus b), \end{aligned} \tag{7}$$

where $\mathcal{F}_{up}(\cdot)$ indicates the bilinear interpolation, $\oplus$ and $\otimes$ express element-wise summation and multiplication.

As presented in Fig. 3, we take the combined feature, i.e., $\mathcal{K}_{3\times3}(\mathcal{F}_{Concat}(a', b'))$ as input, and utilize 1×1 and 3×3 convolutions to obtain two offset maps $\Delta^a, \Delta^b \in \mathbb{R}^{2\times2H\times2W}$. After capturing the offset maps, the final aligned aggregated feature could be performed as

$$out = \mathcal{K}_{3\times3}(\mathcal{F}_A(\mathcal{F}_{up}(a), \Delta^a) \oplus \mathcal{F}_A(b, \Delta^b)), \tag{8}$$

where $\mathcal{F}_A(\cdot,\cdot)$ is the alignment function. Assume that the feature to be aligned is $I \in \mathbb{R}^{C\times H\times W}$, and the non-parameter alignment function is expressed as

$$\begin{aligned} \mathcal{F}_A(I, \Delta) = \sum_{h'}^{H} \sum_{w'}^{W} I_{h',w'} \cdot max(0, 1 - |h + \Delta^1_{h,w} - h'|) \\ \cdot max(0, 1 - |w + \Delta^2_{h,w} - w'|). \end{aligned} \tag{9}$$

Here, h' and w' denote the spatial index of I. The alignment function utilizes the bilinear interpolation kernel on spatial position $(h + \Delta^1_{h,w},\ w + \Delta^2_{h,w})$ to resample the feature in a differentiable manner, where $\Delta^1_{h,w}$, $\Delta^2_{h,w}$ denote the learnable 2-D transformation offsets for pixel $I_{h,w}$. More details on how to find partial derivatives follows [65].

Unlike AlignFA [61], AFAM fully considers the common elements of adjacent features and integrates them by a differentiable learning strategy. Its fusion results are better than the usual functions such as channel-wise concatenation, element-wise summation/multiplication as revealed in Section IV-D.

C. Gated Fold Atrous Spatial Pyramid Pooling (GF-ASPP)

Salient objects in optical RSIs always have diverse scales, which has limited the representation performance of CNNs. ASPP [66] has been widely used for multiscale modeling in CNNs. However, it ignores the relations between different parallel convolution paths, and loses some local correlations due to dilation, which limit its performance on RSI-SOD. We introduce the gate mechanism and unfolding operation [29] into it to explore the intrinsic relevance of adjacent paths and present an enhanced variant, named GF-ASPP.

As presented in Figs. 2 and 4, GF-ASPP is equipped on the end of the hybrid encoder, which consists of four convolutional layers with dilation rates [1,2,4,6] and an image pooling to capture multiple context. Different from ASPP, we apply unfolding operations to merge more context and preserve more local correlation before parallel convolution paths, and the pre-processing of multi-dilation paths can be defined as

$$\widetilde{Y} = \mathcal{F}_{uf}(\mathcal{K}_{3\times3}(S_5)), \tag{10}$$

where $\mathcal{F}_{uf}(\cdot)$ denotes the operation of unfolding, and $\mathcal{K}_{3\times3}(\cdot)$ is utilized to convert S_5 into 128 channels. Assuming $X \in \mathbb{R}^{C\times H\times W}$, the unfolding function uses a $C \times 2 \times 2$ feature window with stride of 2 to slide, and expands the features in the window into a vector of dimensions $4C$ by the channel direction, so the size of X is $4C \times \frac{H}{2} \times \frac{W}{2}$ after unfolding transformation. With this operation, a pixel on the new feature map corresponds to a 2×2 region on the original feature. We can define the reverse operation of unfolding, i.e. folding, which can restore a 4×1×1 map to a 1×2×2 window feature.

In addition, we propose a layer-by-layer guided context learning manner based on gate mechanisms among adjacent dilated convolutional paths. Specifically, the path with a large dilation rate not only receives $\widetilde{Y}$ as input, but also uses the context output of the adjacent convolution path whose dilation rate is only smaller than it as a prior input, which can better preserve the structural knowledge of the image. Furthermore, the gate mechanism performs adaptive spatial attention modeling on the prior input of adjacent layers, eliminates redundant features and refines the prior to assist the path with larger dilation rates to extract more abstract contextual information. The above can be defined as

$$\begin{aligned} Y_1 &= \mathcal{K}_{3\times3}(\widetilde{Y}), \\ Y_i &= \mathcal{K}^{d=2(i-1)}_{3\times3}(\widetilde{Y} \oplus \mathcal{F}_G(Y_{i-1})), \quad i = 2, 3, 4 \\ Y_5 &= \mathcal{F}_{Avgpool}(\widetilde{Y} \oplus \mathcal{F}_G(Y_4)), \end{aligned} \tag{11}$$

where $\mathcal{K}^{d=k}_{3\times3}(\cdot)$ represents a 3×3 convolutional layer with the dilation rate of k. $\mathcal{F}_{Avgpool}(\cdot)$ indicates the image pooling, and $\mathcal{F}_G(\cdot)$ denotes the gate mechanism, i.e.,

$$\mathcal{F}_G(X) = X \otimes \text{Sigmoid}(\mathcal{K}_{3\times3}(X)). \tag{12}$$

After obtaining $Y_1 \sim Y_5 \in \mathbb{R}^{128\times7\times7}$, we restore the output feature to the original spatial dimensions by applying channel concatenation along with folding and a series of convolution operations as shown in Fig. 4, which can be defined as

$$S_6 = \mathcal{K}_{1\times1}(\mathcal{F}_{Fold}(\mathcal{K}_{3\times3}(\mathcal{F}_{cat}(Y_1, Y_2, Y_3, Y_4, Y_5)))). \tag{13}$$

Based on ASPP, we adopt unfolding functions and a layer-by-layer guided learning mechanism to facilitate deep multiscale context extraction, providing the fault-tolerance ability for the operations of the decoder. These improvements are more applicable to optical RSIs with various scales of salient objects.

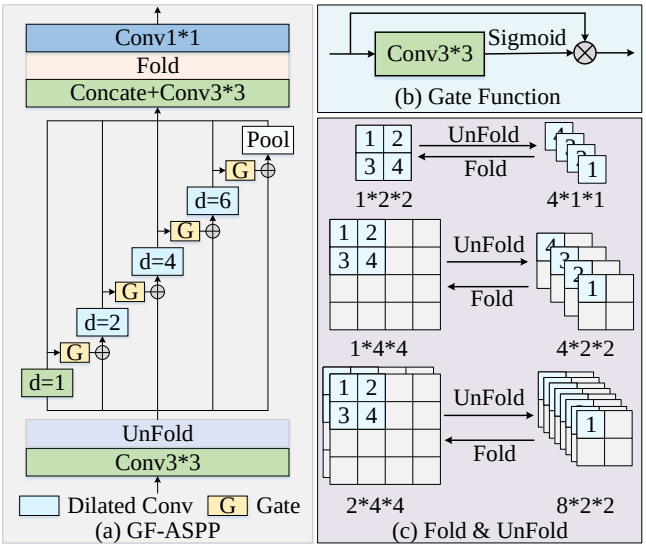


Fig. 4. The framework of the proposed GF-ASPP.

D. Interactive Guidance Loss (IGLoss)

Salient objects in RSIs suffer from complex edges and irregular topology, and existing methods always perform equal supervision on all pixels through cross-entropy loss, which lacks effective usage of edge cues. Recent methods [14], [36] learn saliency maps and boundary auxiliary maps simultaneously, but they ignore the relation between edges and saliency maps. Recent objective functions [38], [39], [67] do not apply the multi-task learning strategy to predict edge auxiliary maps, and cannot realize the intrinsic relation between the two sub-tasks. We seek a dynamic mechanism that allows the network to consider saliency and boundary prediction simultaneously, and perform mutual supervision in the training process.

How to mutually supervise saliency map and edge auxiliary map in a dynamical manner? We associate that the outputs of the two branches of the saliency detection head (as illustrated in Fig. 2(b)) are adaptively adjusted. Therefore, our goal is to utilize this knowledge to enhance the traditional cross-entropy loss. To this end, we propose to not only jointly model the edge prediction and saliency detection tasks, but also make them mutually reinforcing to each other's performance. Specifically, we achieve this by considering two aspects. One is that if a pixel is predicted as an edge sample, it is then treated as a hard example. Thus, its corresponding weight for saliency detection loss should be increased. The other is that if an edge pixel is correctly predicted as a true positive sample, then it can be viewed as an easy sample. Hence, its corresponding balancing weight for edge modeling loss should be decreased. Formally, the proposed IGLoss is defined as:

$$\mathcal{L}_{IG}(P^s, P^e, G^s, G^e) = \mathcal{L}_e(P^e, P^s, G^e) + \mathcal{L}_s(P^s, P^e, G^s), \quad (14)$$

$$\mathcal{L}_e(P^e, P^s, G^e) = \frac{1}{n}\sum_{i=1}^{n}((1 - p_i^s g_i^e)m_2 + 1) l_{bce}(p_i^e, g_i^e), \quad (15)$$

$$\mathcal{L}_s(P^s, P^e, G^s) = \frac{1}{n}\sum_{i=1}^{n}(p_i^e m_1 + 1) l_{bce}(p_i^s, g_i^s), \quad (16)$$

$$l_{bce}(x, y) = -ylog(x) - (1 - y)log(1 - x), \quad (17)$$

where P^s, P^e, G^s, and G^e express the predicted saliency map, the predicted edge map, ground truth saliency map and ground truth edge map. p_i^s, p_i^e, g_i^s and g_i^e are the pixels in P^s, P^e, G^s, and G^e, respectively. i denotes the index of each pixel while n represents the total number of pixels. And m_1 and m_2 are the weight factors to balance the hard and easy examples.

IGLoss adaptively facilitates the learning process of the two sub-tasks by resampling the loss in the boundary areas based on the dynamic features predicted by the model. Obviously, for objects with blurred edges and complex topology, the model trained with IGLoss can have better representation ability for the pixels which are close to boundaries in RSIs.

E. Total Loss Function with Deep Supervision

The proposed HFANet employs IGLoss and weighted Intersection Over Union (IoU) as loss functions to train the network with deep supervision (as illustrated in Fig. 2(a)). For the detection stage i, the loss function is defined as:

$$\mathcal{L}_i = \lambda_1 \mathcal{L}_{IG}(P^s, P^e, G^s, G^e) + \lambda_2 \mathcal{L}_{wiou}(P^s, G^s), \quad (18)$$

where weighted IoU loss is defined as:

$$\mathcal{L}_{wiou}(P^s, G^s) = 1 - \frac{\sum_1^n \text{mul}(p_j^s, g_j^s) + 1}{\sum_1^n (\text{sum}(p_j^s, g_j^s) - \text{mul}(p_j^s, g_j^s)) + 1}, \quad (19)$$

where $\text{sum}(p_j^s, g_j^s)$, $\text{mul}(p_j^s, g_j^s)$ indicate the sum, and production of the predicted saliency map and its ground-truth at pixel j, respectively.

To better balance the five detection stages of the model, inspired by SARNet [35], we perform a weighted summation of the losses of the five stages, i.e.

$$\mathcal{L}_{total} = \mathcal{L}_1 + \sum_{i=2}^{5} \frac{1}{2^{i-2}} \mathcal{L}_i. \quad (20)$$

Empirically, we find that setting the hyperparameters λ_1 and λ_2 in $\mathcal{L}_i$ are all set to 1, and m_1, m_2 in $\mathcal{L}_i$ are equal to 4 work well in our experiments.

IV. Experiments

A. Experimental Protocol

1) Datasets: We perform experiments on three public optical RSI-SOD datasets with pixel-wise annotations.

ORSSD [12] contains 800 optical RSIs with various salient object diversity and cluttered background. We employ a consistent strategy with that in [12], using 600 images to train and the rest 200 images to test.

EORSSD [13] is an extension of ORSSD, which includes 2000 more comprehensive optical RSIs with more abundant scenarios. It consists of 1400 images as the training subset and the other 600 images for testing.

ORSI-4199 [14] is the latest and most challenging optical RSI-SOD dataset. This dataset is divided into 2000 images for training and the remainder 2199 ones as testing subset. Besides, it defines nine different scene attributes, which helps us objectively evaluate SOD models under various attributes.

2) Evaluation Metrics: The broadly-used criteria, MAE, F-measure, and S-measure are applied for quantitative evaluation. Besides, the structure similarity score (SSIM) [68] is utilized to measure attributes-based performance, and we draw the PR and F-measure curves for a more clear comparison.

PR curve: [3] Suppose SM, GT as the predicted saliency map and ground truth, we can obtain SM and GT as binary masks by using the thresholds ranging from [0, 255]. Then, the combinations of precision and recall under different thresholds could be further estimated as follows:

$$\text{Precision} = \frac{|\text{SM} \cap \text{GT}|}{|\text{SM}|}, \quad \text{Recall} = \frac{|\text{SM} \cap \text{GT}|}{|\text{GT}|}. \tag{21}$$

We can draw PR curve based on 256 pairs of precision and recall values. The closer PR curve is to the upper right, the better the performance of the model is.

MAE [45] defines the pixel-wise difference between SM and GT as follows

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\text{SM}(i) - \text{GT}(i)|, \tag{22}$$

where n indicates the total number of pixels. Note that SM is a continuous map from [0,1] and GT is a binary map $\in \{0, 1\}$. For a saliency detector, MAE score is smaller, the detection performance is better, obviously.

F-measure [3] is defined as the weighted combination of precision and recall value for saliency maps, i.e.,

$$F_\beta = \frac{(1+\beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}, \tag{23}$$

where β^2 is set to 0.3 for emphasizing the precision over recall as recommended in [3]. We report the maximum F-measure under different thresholds from 0 to 255 and draw F-measure curves simultaneously. Generally, a more expressive model can achieve a higher max F-measure score and a larger coordinate area covered by F-measure curve.

S-measure [46] applies the structural similarity of region-aware (S_r) and object-aware (S_o) levels to evaluate the structural information of saliency maps, which is defined as

$$S_m = \alpha \times S_o(\text{SM}, \text{GT}) + (1-\alpha) \times S_r(\text{SM}, \text{GT}), \tag{24}$$

where α is equal to 0.5 for balancing the two structural similarities as suggested in [46].

3) Implementation Details: We train and test all methods on the three RSI-SOD datasets, respectively. All experiments are implemented on the PyTorch1.8 toolbox in Ubuntu 18.04 operating system with an NVIDIA GeForce RTX 3090 GPU. For a fair comparison, we retrain and test all deep learning-based approaches [13], [14], [22]–[31], [33]–[36] by the official public source codes with the default parameter settings.

Note that for all algorithms, the input images are uniformly resized to 448×448 for training and testing, and the batch size is unified as 8 during training. We adopt flipping and rotation as data augmentation as recommended in [12], [13], [34] and acquire 7 additional augmented versions of the original images for all algorithms. Our network loads the pretrained weights of the hybrid encoder for training with the Adamw optimizer and Cosine learning rate scheduler, and other convolutional layers are initialized using *kaiming-normal* strategy. The model is trained for 300 epochs, and the initial learning rate, momentum and weight decay are set to 5e-4, 0.9, 0.05, respectively. The codes, models and saliency maps of all methods are available for facilitating future research.

B. Comparison With State-of-the-Art Methods

We compare our HFANet with 18 state-of-the-art models on three datasets, of which six ones for optical RSIs (i.e., SARNet [35], DAFNet [13], MCCNet [33], MJRBM [28], RRNet [34], EMFINet [36]), two traditional models for NSIs (i.e., LC [2] and FT [3]), and ten CNN-based approaches for NSIs (i.e., NLDF [22], DSS [23], RAS [24], PoolNet [25], PFAN [26], MINet [27], SCRN [28], GateNet [29], F3Net [30], PFSNet [31]). In what follows, we present the quantitative comparison, qualitative comparison, attribute-based study and computational complexity comparison.

1) Quantitative Comparison: To evaluate quantitative comparison on all three datasets, we first draw PR and F-measure curves in Fig. 5. We find that our HFANet-R performs better than other competitors on all three datasets. As displayed in Fig. 5, PR curve of our model is closer to the upper right corner, and the area covered by F-measure curve of ours is also the largest one among all competitors.

For a more intuitive comparison, quantitative results in terms of MAE, max F-measure (F_β) and S-measure (S_m) are reported in Table I. These models can be divided into three types, of which there are 2 traditional algorithms, 12 VGG-based models and 8 ResNet-based models. Among them, the performance of deep learning-based methods such as NLDF [22] and DSS [23] is more significant than traditional approaches [2], [3]. Besides, among the algorithms based on VGG or ResNet, most RSI-SOD algorithms also outperform NSI-SOD models. Typically, MCCNet [33] acquires better performance compared with VGG-based NSI-SOD methods on three datasets. This justifies the necessity of designing saliency models for optical RSIs exclusively. It can also be concluded from Table I that HFANet-V and HFANet-R achieve the best performance on almost all evaluation metrics among competitors. The proposed HFANet-V even reaches comparable results to EMFINet-V [36] which adopts the image pyramid strategy. Compared with the state-of-the-art RSI-SOD models [14], [34], [36], HFANet-R outperforms them by a large margin, especially on the ORSI-4199 dataset, and it obtains a superior performance of 0.0329 in terms of MAE. It performs best on all three datasets, and provides a state-of-the-art comparative benchmark for RSI-SOD. We attribute this performance gain to the global context extraction and feature alignment better than other models, and this is why our model is more favorable for optical RSIs. To reveal the benefits brought by each module, we conduct more detailed analysis in Section IV-C and IV-D.

2) Qualitative Comparison: To qualitatively compare all approaches, typical visualized saliency maps on the ORSI-4199 are shown in Fig. 6. For each test image, we provide the original image, ground-truth and the prediction results of 19 algorithms, among which MJRBM [14], EMFINet [36]

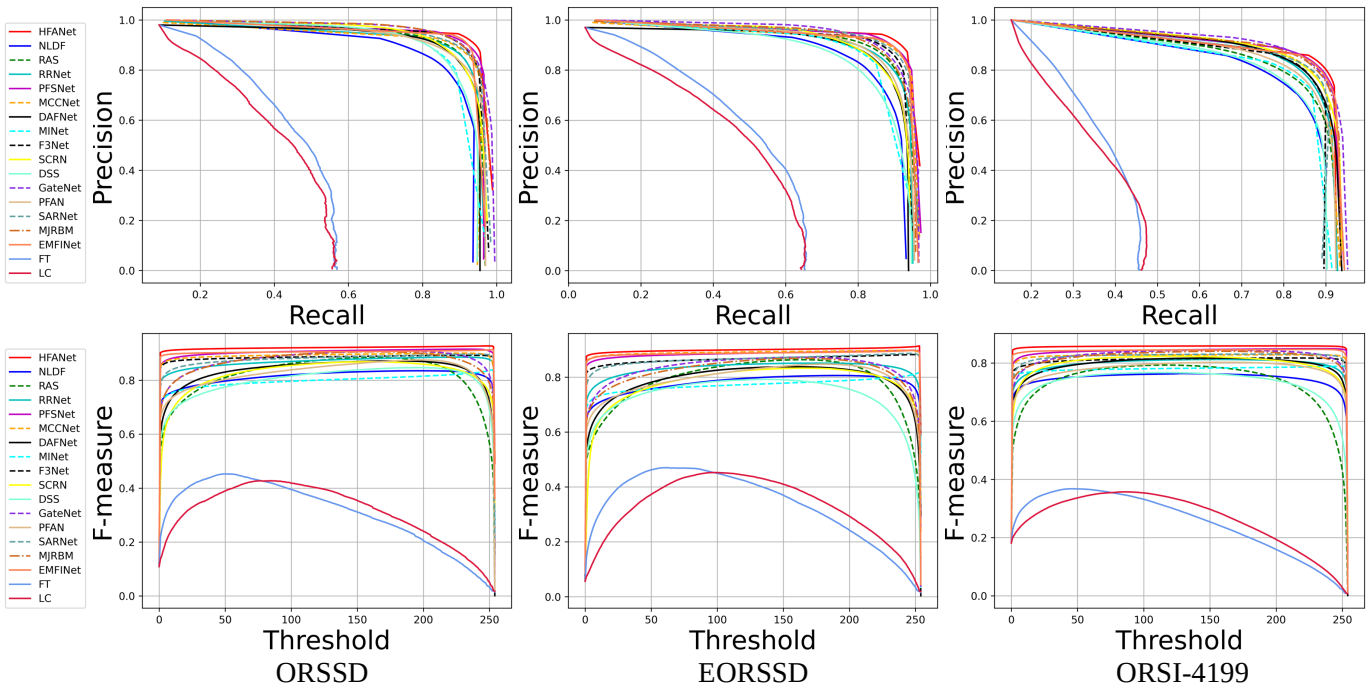


Fig. 5. Quantitative comparisons of the presented model with 17 state-of-the-art methods. The first and second rows are PR and F-measure curves, respectively.

and our HFANet only utilizes ResNet versions. Intuitively, the predicted results of HFANet are more complete, precise, and closer to ground-truths than other methods in Fig. 6(u). Specifically, our algorithm has the following three advantages:

a) Superiority in scenes with complex background or low-contrast: As shown in Fig. 6, traditional models cannot detect salient objects effectively with complex background or low contrast. Among competitors, HFANet does not recognize background regions as salient objects. In addition, for low-contrast saliency objects, many CNN-based methods recognize the background as foreground, and even predict cluttered saliency maps (e.g., the last three lines in Figs. 6(i)-(s)), while our method can well deal with the interference of this problem. The above reveals that the adaptive local/global features extracted by the hybrid encoder can be successfully applied to detect complex background and low-contrast scenes.

b) Superiority in scenes with multiple or multiscale objects: Whether there are large, small, tiny, or multiple objects in a RSI, our model can detect the complete salient regions, and introduce minimal background confusion. Typically, some CNN-based methods [22]–[31] perform even worse than traditional models in rows 10-12 of Figs. 6(e)-(q), which identify the background as salient regions or cannot separate multiple salient objects clearly. It is obvious that HFANet performs well in these scenarios benefiting from the superiority of multiscale context modeling and feature alignment.

c) Superiority in salient regions with complicated edges or irregular topology: The island, river and building in rows 4-6 of Fig. 6 are the most challenging examples of RSI-SOD, and some algorithms fail in these scenes due to complex boundaries and irregular topology. To detect complete saliency regions, we propose IGloss to promote the learning ability for irregular objects through the mutual guidance of regions and edges. By contrast, our model not only outperforms MJRBM [14], EMFINet [36], and SARNet [35] utilizing edge learning modules, but also does not introduce any parameters.

3) Attribute-Based Study: Each image in the ORSI-4199 dataset is described as representing one or more attributes of common challenges in RSIs. These annotations help researchers discover the pros and cons of SOD methods. Table IV reveals SSIM scores of the presented HFANet and 15 state-of-the-art methods. We find that our method ranks in the top three for five attributes among these nine. In addition, it ranks best in total average score. The above denotes that HFANet expresses other models in the majority of various scenes with balanced scores. Although EMFINet [36], SARNet [35], GateNet [29] and PFSNet [31] obtain the highest scores in several scenarios, they always perform very poorly on other attributes simultaneously.

4) Computational Complexity Comparison: We utilize four metrics, i.e., floating point operations (FLOPs), number of parameters, memory footprint, and inference speed to measure the model complexity, and report them in Table II. First of all, our models have the fewest FLOPs among competitors, which shows a great computational advantage. In terms of memory footprint and model parameters, our models are in the midstream level. Compared with RRNet [34] and EMFINet [36] with heavy FLOPs and parameters, HFANet-R performs worse in inference speed but is superior in the other three indicators. We speculate that the global operations introduced by the hybrid encoder largely limits the inference speed. To prove this, two different versions of HFANet with hybrid encoder and VGG16 are compared in Table II. By contrast, we observe that the inference speed of HFANet-VGG16 has been greatly improved (from 36.18 FPS to 44.87 FPS). Note that the increase of FLOPs and parameters is mainly due to

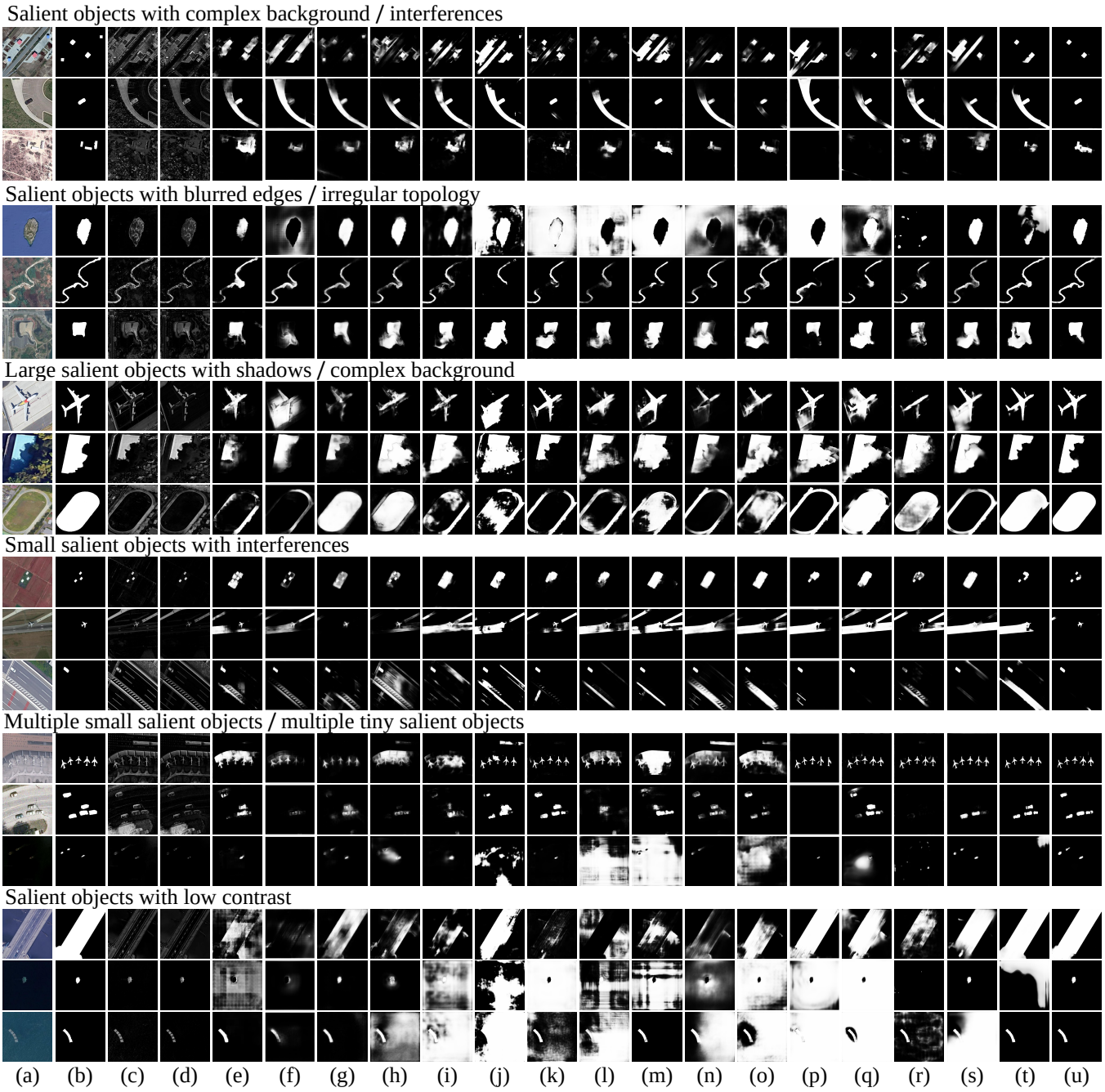


Fig. 6. Predicted results comparisons with 18 state-of-the-art approaches on the challenging ORSI-4199 dataset [14], including 6 CNN-based RSI-SOD methods, 10 CNN-based NSI-SOD methods, and 2 traditional NSI-SOD methods, on various patterns. Please zoom-in for the best view. (a) Optical RSIs. (b) GT. (c) LC [2]. (d) FT [3]. (e) NLDF [22]. (f) DSS [23]. (g) RAS [24]. (h) PoolNet [25]. (i) PFAN [26]. (j) MINet [27]. (k) SARNet [35]. (l) DAFNet [13]. (m) MCCNet [33]. (n) SCRN [28]. (o) GateNet [29]. (p) F3Net [30]. (q) PFSNet [31]. (r) RRNet [34]. (s) MJRBM-R [14]. (t) EMFINet-R [36]. (u) Ours-R.

changes in the fourth AFAM module. The above shows that the hybrid encoder really limits the inference speed because of global operations introduced by Transformer blocks. Thus how to design a lightweight and efficient SOD model for optical RSIs will be further investigated in our future study.

C. Ablation Study

We perform comprehensive ablation experiments on the EORSSD dataset to reveal the effects of our proposed modules and loss function. We find consistently the increasing trends of detection performance on the EORSSD dataset based on these quantitative results in Table III. Besides, some visual examples shown in Fig. 7 also prove the effectiveness of the modules and objective function designed in our network intuitively.

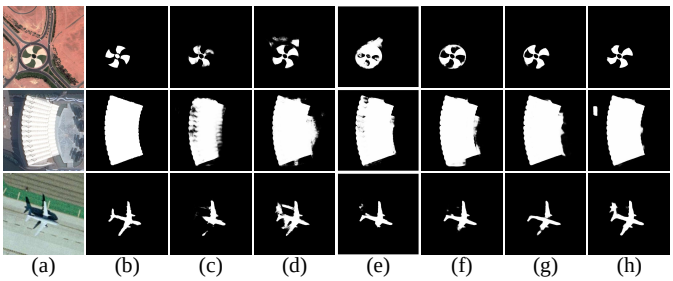
1) Baseline Setup: The encoder of the baseline is ResNet50 network, and the decoder adopts feature pyramid architecture, in which the decoder utilizes channel-wise concatenation for feature fusion. The baseline only applies the cross-entropy loss function at the shallowest layer of the decoder to supervise the edge and saliency map respectively. In all ablation exper-

TABLE I
QUANTITATIVE RESULTS ON THREE OPTICAL RSI-SOD DATASETS. THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN AND BLUE, RESPECTIVELY.

Methods	Publication	Backbone	Input Size	ORSSD Dataset [12]			EORSSD Dataset [13]			ORSI-4199 Dataset [14]		
				$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$
Traditional Methods												
LC [2]	MM 2006	–	448×448	0.4275	0.1230	0.5941	0.4526	0.0864	0.5954	0.3573	0.1893	0.5270
FT [3]	CVPR 2009	–	448×448	0.4526	0.1126	0.5916	0.4704	0.0715	0.6107	0.3680	0.1791	0.5256
VGG-based												
NLDF [22]	CVPR 2017	VGG16	448×448	0.8352	0.0267	0.8702	0.8060	0.0154	0.8706	0.7639	0.0584	0.8053
DSS [23]	CVPR 2017	VGG16	448×448	0.8469	0.0268	0.8688	0.7921	0.0167	0.8371	0.7672	0.0561	0.8115
RAS [24]	ECCV 2018	VGG16	448×448	0.8841	0.0185	0.8896	0.8636	0.0114	0.8800	0.7930	0.0595	0.8142
PoolNet [25]	CVPR 2019	VGG16	448×448	0.8291	0.0268	0.8610	0.8121	0.0207	0.8279	0.7777	0.0573	0.8184
PFAN [26]	CVPR 2019	VGG16	448×448	0.8755	0.0207	0.8853	0.8472	0.0127	0.8848	0.8024	0.0486	0.8373
MINet [27]	CVPR 2020	VGG16	448×448	0.8380	0.0227	0.8640	0.8178	0.0129	0.8569	0.7891	0.0473	0.8232
SARNet [35]	RS 2021	VGG16	448×448	0.8963	0.0185	0.8976	0.8865	0.0102	0.9097	0.8309	0.0448	0.8536
DAFNet [13]	TIP 2021	VGG16	448×448	0.8717	0.0161	0.8982	0.8378	0.0106	0.8824	0.8169	0.0473	0.8477
MCCNet [33]	TGRS 2022	VGG16	448×448	0.9005	0.0212	0.9040	0.8976	0.0083	0.9306	0.8284	0.0439	0.8506
MJRBM-V [14]	TGRS 2022	VGG16	448×448	0.9028	0.0140	0.9156	0.8705	0.0099	0.9088	0.8352	0.0392	0.8601
EMFINet-V [36]	TGRS 2022	VGG16	448×448	0.9224	0.0112	0.9323	0.8931	0.0083	0.9240	0.8438	0.0373	0.8648
HFANet (Ours)	–	Hybrid	448×448	0.9224	0.0113	0.9324	0.9007	0.0082	0.9292	0.8419	0.0379	0.8659
ResNet-based												
SCRN [28]	ICCV 2019	ResNet50	448×448	0.8687	0.0210	0.8799	0.8326	0.0158	0.8288	0.8232	0.0423	0.8524
GateNet [29]	ECCV 2020	ResNet50	448×448	0.9083	0.0125	0.9103	0.8724	0.0091	0.9010	0.8443	0.0387	0.8660
F3Net [30]	AAAI 2020	ResNet50	448×448	0.8927	0.0126	0.9245	0.8822	0.0077	0.9218	0.8175	0.0435	0.8520
PFSNet [31]	AAAI 2021	ResNet50	448×448	0.9153	0.0101	0.9303	0.8979	0.0077	0.9287	0.8496	0.0374	0.8686
RRNet [34]	TGRS 2022	ResNet50	448×448	0.8857	0.0142	0.9110	0.8511	0.0101	0.8964	0.8122	0.0448	0.8449
MJRBM-R [14]	TGRS 2022	ResNet50	448×448	0.9058	0.0129	0.9128	0.8685	0.0092	0.8980	0.8406	0.0379	0.8685
EMFINet-R [36]	TGRS 2022	ResNet34	448×448	0.9132	0.0107	0.9350	0.8972	0.0075	0.9286	0.8469	0.0352	0.8678
HFANet (Ours)	–	Hybrid	448×448	0.9274	0.0090	0.9376	0.9138	0.0069	0.9363	0.8590	0.0329	0.8758

TABLE II
MODEL COMPLEXITY COMPARISON ON THE ORSI-4199 DATASET.

Methods	FLOPs (G)	Params (M)	Memory (M)	FPS
MINet [27]	287.05	47.56	3881	40.62
DAFNet [13]	839.21	29.35	4753	58.52
MCCNet [33]	358.64	67.65	4515	51.49
EMFINet-R [36]	541.05	95.09	7373	27.75
MJRBM-V [14]	155.09	43.78	1849	29.21
MJRBM-R [14]	80.56	63.28	2313	21.79
PFSNet [31]	73.55	31.18	1799	22.10
RRNet [34]	692.15	86.27	5675	41.54
HFANet-VGG16	87.51	49.27	2313	44.87
HFANet-V	64.63	43.59	2365	36.18
HFANet-R	68.32	60.53	2667	26.28



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 7. Quantitative comparisons of ablation study for our model. (a) Optical RSIs. (b) GT. (c) baseline. (d) baseline+hybrid encoder. (e) baseline+hybrid encoder+GF-ASPP. (f) baseline+hybrid encoder+GF-ASPP+AFAM. (g) baseline+hybrid encoder+GFASPP+AFAM+DS. (h) the proposed HFANet.

TABLE III
ABLATION EXPERIMENTS ON THE EORSSD [13] DATASET. HE (HYBRID ENCODER), GFA (GF-ASPP), AFAM, DS (DEEP SUPERVISION) AND IGL (IGLOSS). THE BEST RESULTS ARE MARKED IN BOLD.

No.	HE	GFA	AFAM	DS	IGL	MAE$\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$
1						.0094	.8768	.9129
2	✓					.0082	.8952	.9233
3		✓				.0089	.8794	.9113
4			✓			.0080	.8810	.9188
5				✓		.0087	.8868	.9163
6					✓	.0094	.8794	.9146
7	✓	✓				.0073	.8990	.9298
8		✓		✓		.0076	.8918	.9225
9	✓	✓	✓			.0071	.9053	.9325
10		✓	✓	✓		.0076	.8940	.9265
11		✓	✓	✓	✓	.0075	.8965	.9262
12	✓	✓	✓	✓		.0069	.9072	.9330
13	✓	✓	✓	✓	✓	**.0069**	**.9138**	**.9363**

iments, all hyperparameters of model are consistent for fair comparison. Baseline reaches 0.0094 MAE, 0.8768 F_β, and 0.9129 S_m on the EORSSD dataset.

2) Effects of Hybrid Encoder: We propose the hybrid encoder to adaptively extract the global/local context of optical RSIs, thereby mitigating the influence of complex background. Comparing with rows 1, 2 and 10, 12 in Table III respectively, the results can illustrate the positive significance of the hybrid encoder. When integrating the hybrid encoder into the baseline, the model increases F_β from 0.8768 to 0.8952 and S_m from 0.9129 to 0.9233, and it illustrates the effectiveness of the hybrid encoder. When removing the hybrid encoder of the complete HFANet, F_β and S_m become smaller and MAE increases. The above demonstrates that the hybrid encoder is superior to pure CNN-based encoder. For more data and visual analysis of the hybrid encoder, see Section IV-D.

3) Effects of AFAM: To handle the feature misalignment issue, we present AFAM to obtain powerful multiscale spatial

TABLE IV
ATTRIBUTE-BASED TEST RESULTS ON THE ORSI-4199 DATASET [14]. WE PRESENT THE AVERAGE STRUCTURE SIMILARITY SCORE FOR ALL TEST IMAGES WITH THAT PARTICULAR ATTRIBUTES AS [14]. THE LAST ROW SHOWS THE AVERAGE PERFORMANCE. THESE TOP THREE SCORES ARE HIGHLIGHTED IN RED, GREEN AND BLUE, RESPECTIVELY.

Attr	NLDF	DSS	RAS	PFAN	MINet	SARN	DAFN	MCCN	SCRN	GateN	F3Net	PFSN	RRNet	MJRBM	EMFI	Ours
BSO	.7365	.7604	.7472	.7984	.7929	.7984	.8064	.8200	.8295	.8546	.8256	.8531	.8226	.8344	.8547	.8465
CS	.7395	.7512	.7650	.8021	.7858	.8123	.8151	.8210	.8274	.8539	.8234	.8519	.8138	.8451	.8509	.8476
CSO	.7074	.7368	.7313	.7714	.7630	.7665	.7831	.7935	.8001	.8234	.7887	.8196	.7900	.8003	.8250	.8158
ISO	.6496	.7088	.6595	.7280	.7184	.7573	.7485	.7778	.7762	.8110	.7872	.8166	.7538	.7868	.8102	.8102
LCS	.6305	.6369	.6816	.6756	.6649	.7182	.7018	.7119	.7094	.7318	.7071	.7342	.6886	.7332	.7225	.7429
MSO	.7296	.7266	.7234	.7739	.7429	.8162	.7817	.7720	.7949	.8156	.7902	.8218	.7702	.8183	.7905	.8081
NSO	.6719	.6674	.7356	.7496	.7304	.7949	.8054	.8028	.7933	.8552	.7916	.8318	.7873	.8391	.8412	.8629
OC	.7085	.7131	.6915	.7457	.6707	.8071	.7780	.7721	.7474	.7730	.7481	.7861	.7301	.8055	.7662	.8039
SSO	.6632	.6585	.6703	.6984	.6447	.7596	.7219	.7094	.7092	.7311	.7179	.7493	.6897	.7456	.7130	.7495
Avg	.6930	.7067	.7117	.7492	.7237	.7812	.7713	.7756	.7764	.8055	.7755	.8072	.7607	.8009	.7971	.8097

features adaptively. It is shown that AFAM improves the detection performance by rows 1, 4 or 7, 9 in Table III. In particular, by assembling AFAM modules into the baseline, the MAE indicator is significantly reduced, from 0.0094 to 0.0080. It reveals that the fusion and alignment strategies of spatial features via AFAM really contributes to the model for interpreting multiscale salient objects in optical RSIs.

4) Effects of GF-ASPP: Considering that capturing multi-scale semantic information is beneficial to detect salient objects with imbalanced scale variation, and the shortcomings of ASPP in design, we propose GF-ASPP. When only integrating GF-ASPP into the baseline, it can be seen in rows 1, 3 of Table III that the boost of model performance is not obvious. This is because the abundant feature extracted by GF-ASPP is greatly reduced after multiple propagation in the baseline (without deep supervision (DS)). When our method merges GF-ASPP and DS together, we observe that the model performance is further improved (1.5% $F_\beta\uparrow$ especially), which indicates the combination of GF-ASPP and DS can reach better SOD results in optical RSIs.

5) Effects of IGLoss: To detect salient objects with complex edges and irregular topology effectively, we propose IGLoss to jointly learn salient regions and their edges. The effectiveness of IGLoss can be illustrated by rows 1, 6 or 10, 12 of Table III. Compared with baseline and baseline+IGLoss, the model performance only has a slight improvement in the two indicators F_β and S_m. But comparing HFANet trained with IGLoss and HFANet trained without IGLoss, we observe a large reduction on F_β (6.6%$\downarrow$). The above shows the incremental contribution of IGLoss to our approach. For more comparison and visual analysis, see Section IV-D.

D. Model Analysis

We set up more adequate experiments and visual analysis to compare and illustrate the proposed key modules and objective function in this subsection.

1) Analysis of Hybrid Encoder: To reveal why our hybrid encoder can boost saliency detection performance, we set up several different encoders to build HFANet, including VGG16, hybrid VGG16, ResNet50, hybrid ResNet50, and conduct comparison experiments on all three datasets. First, the quantitative results in Table V clearly show that both hybrid encoders outperform pure CNN-based encoders on all datasets. Second, the feature visualizations in Fig. 8 can illustrate the following three advantages of the hybrid encoder:

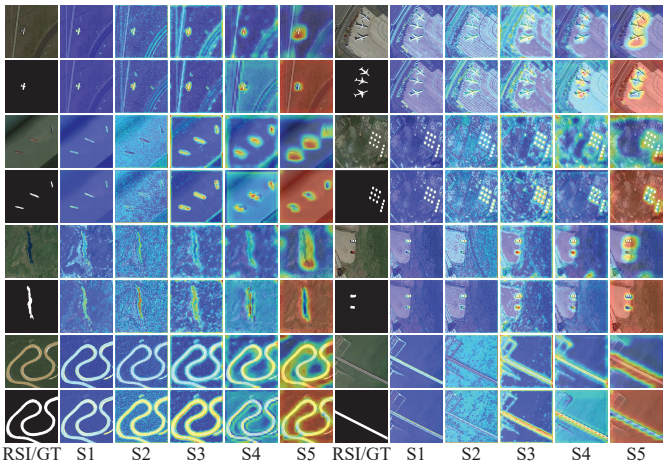


Fig. 8. Feature visualizations comparisons of the proposed Hybrid Encoder with ResNet50. The odd and even rows are average encoder features (S1-S5) of ResNet50 and Hybrid Encoder, respectively.

a) Extracting global context information in S5: The hybrid encoder introduces PVT blocks in the last two stages, the purpose of which is to explore global context information in optical RSIs. By comparing the odd and even rows of S5 in Fig. 8, we find the peculiarity of the hybrid encoder, i.e., it can accurately extract the global context of salient regions, while the CNN-based encoder only obtains the local context roughly. We attribute the performance boost to this difference mainly as illustrated in Table V.

b) Capturing more accurate local context in S4: As shown in Fig. 8, whether it is airplane, ship, road, storage tank, car, the hybrid encoder can detect more accurate and refined local features than ResNet50. Considering both S4 and S5, it can be proved that our hybrid encoder can adaptively focus on global/local information of salient regions in different stages, and it has stronger feature representation ability than ResNet50 for the interference of cluttered background of optical RSIs.

c) Exploring cleaner local context information in S1∼S3: Unexpectedly, the hybrid encoder that introduces the transformer blocks in the deep stages (S4∼S5) can facilitate local semantic understanding for optical RSIs in the shallow CNN-based stages (S1∼S3). Intuitively, the hybrid encoder detects more detailed information in shallow stages and has a larger

TABLE V
COMPARISON EXPERIMENTS OF THE PROPOSED HFANET WITH DIFFERENT ENCODER NETWORKS ON THREE DATASETS [12]–[14].

Encoder Network	ORSSD Dataset [12]			EORSSD Dataset [13]			ORSI-4199 Dataset [14]		
	$F_\beta\uparrow$	MAE↓	$S_m\uparrow$	$F_\beta\uparrow$	MAE↓	$S_m\uparrow$	$F_\beta\uparrow$	MAE↓	$S_m\uparrow$
VGG16	0.9014	0.0140	0.9215	0.8966	0.0083	0.9252	0.8340	0.0413	0.8610
Hybrid VGG16	**0.9224**	**0.0113**	**0.9324**	**0.9007**	**0.0082**	**0.9292**	**0.8419**	**0.0379**	**0.8659**
ResNet50	0.9185	0.0119	0.9325	0.8965	0.0078	0.9252	0.8548	0.0352	0.8681
Hybrid ResNet50	**0.9274**	**0.0090**	**0.9376**	**0.9138**	**0.0069**	**0.9363**	**0.8590**	**0.0329**	**0.8758**

TABLE VI
COMPARISON STUDY OF AFAM WITH TYPICAL AGGREGATION STRATEGIES ON THE ORSI-4199 DATASET.

Feature Aggregation Strategies	MAE↓	$F_\beta\uparrow$	$S_m\uparrow$
Baseline (channel-wise concatenation)	0.0407	0.8310	0.8561
+ element-wise summation	0.0410	0.8244	0.8540
+ element-wise multiplication	0.0397	0.8311	0.8589
+ WCAT [31]	0.0378	**0.8448**	**0.8637**
+ AlignFA [61]	0.0383	0.8357	0.8563
+ proposed AFAM	**0.0368**	0.8399	0.8589

TABLE VII
COMPARISON STUDY OF GF-ASPP WITH TYPICAL CONTEXT METHODS ON THE ORSI-4199 DATASET.

Multiscale Context Methods	MAE↓	$F_\beta\uparrow$	$S_m\uparrow$
Baseline with deep supervision	0.0407	0.8310	0.8561
+ PPM [69]	0.0387	0.8417	0.8588
+ ASPP [66]	0.0381	0.8392	0.8550
+ ASPP + unfolding	0.0380	0.8437	0.8621
+ proposed GF-ASPP	**0.0372**	**0.8468**	**0.8628**

activation response for salient regions. For objects in Fig. 8 such as planes, roads, ships, islands, compared to ResNet50, the hybrid encoder can suppress more background noise in S2, and extract more clear spatial features of foreground objects in S3 simultaneously.

2) Analysis of AFAM: The proposed AFAM outperforms the basic feature aggregation strategies, i.e., element-wise summation/multiplication, channel-wise concatenation, as described in Table VI. It is worth mentioning that AFAM achieves comparable quantitative results with respect to WACT on the ORSI-4199 dataset, where WACT is the state-of-the-art module incorporating the specialized attention mechanism [31]. Unlike WACT, AFAM employs a nearly parameter-free alignment operation to integrate adjacent features. We also conduct a comparison experiment, i.e., baseline+AlignFA, to validate the effectiveness of the proposed AFAM. We observe that the performance of AlignFA [61] really decreases a lot, and is close to the baseline. This is mainly because AlignFA does not pay attention to the common elements between adjacent features. Typical feature visualizations of decoder (out1∼out4) by baseline+AFAM and baseline are shown in Fig. 9. We clearly discover that the feature maps obtained by AFAM focus more on salient regions and have less background interference. In addition, there is the most positive effect in out2, i.e., features guided by multiple AFAMs contain almost no background noise of optical RSIs.

3) Analysis of GF-ASPP: The presented GF-ASPP integrates gate strategies, unfolding mechanisms and spatial pyramid pooling layers to extract multiscale semantic information from optical RSIs. To verify the superiority of GF-ASPP, we select two mainstream modules, namely, PPM [69], ASPP [66], and the results on the ORSI-4199 dataset are as shown in Table VII. We find that all modules exceed the baseline, revealing the importance of multiscale context modeling for RSI-SOD. Besides, GF-ASPP takes the output of low dilation layers as the prior input of higher dilation layers, and this mechanism of guiding information flow layer by layer makes it perform better than competitors. To explore the effects of gate and unfolding mechanism, we add a variant named "ASPP+unfolding", i.e., GF-ASPP without gates. First, compared with GF-ASPP and ASPP+unfolding, the former has a greater advantage, especially on F_β reaching 0.8468. It shows that the gate strategy provides a selection mechanism as information transmission between adjacent layers.Second, ASPP+unfolding has some advantages over ASPP in all metrics, which shows the effectiveness of the unfolding operations.

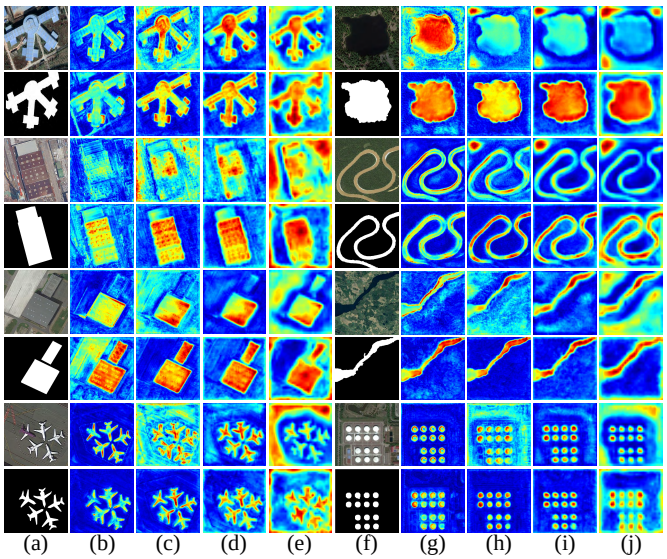


Fig. 9. Features visualizations of the AFAM. The odd and even rows are decoder features (out1-out4) of baseline and baseline+AFAM, respectively. (a), (f) RSIs or GTs. (b), (g) Out1. (c), (h) Out2. (d), (i) Out3. (e), (j) Out4.

4) Analysis of IGLoss: To reveal the superiority of IGLoss, we select the recent state-of-the-art objective functions [37]–[39] for comparison. Among them, F-measure Loss supervises the SOD model by maximizing F_β metric, and CT and ACT functions calculate the final loss by increasing the weight of the pixels in the boundary area through fixed threshold or adaptive strategy. However, these objective functions do not take into account the correlation between the two sub-tasks,

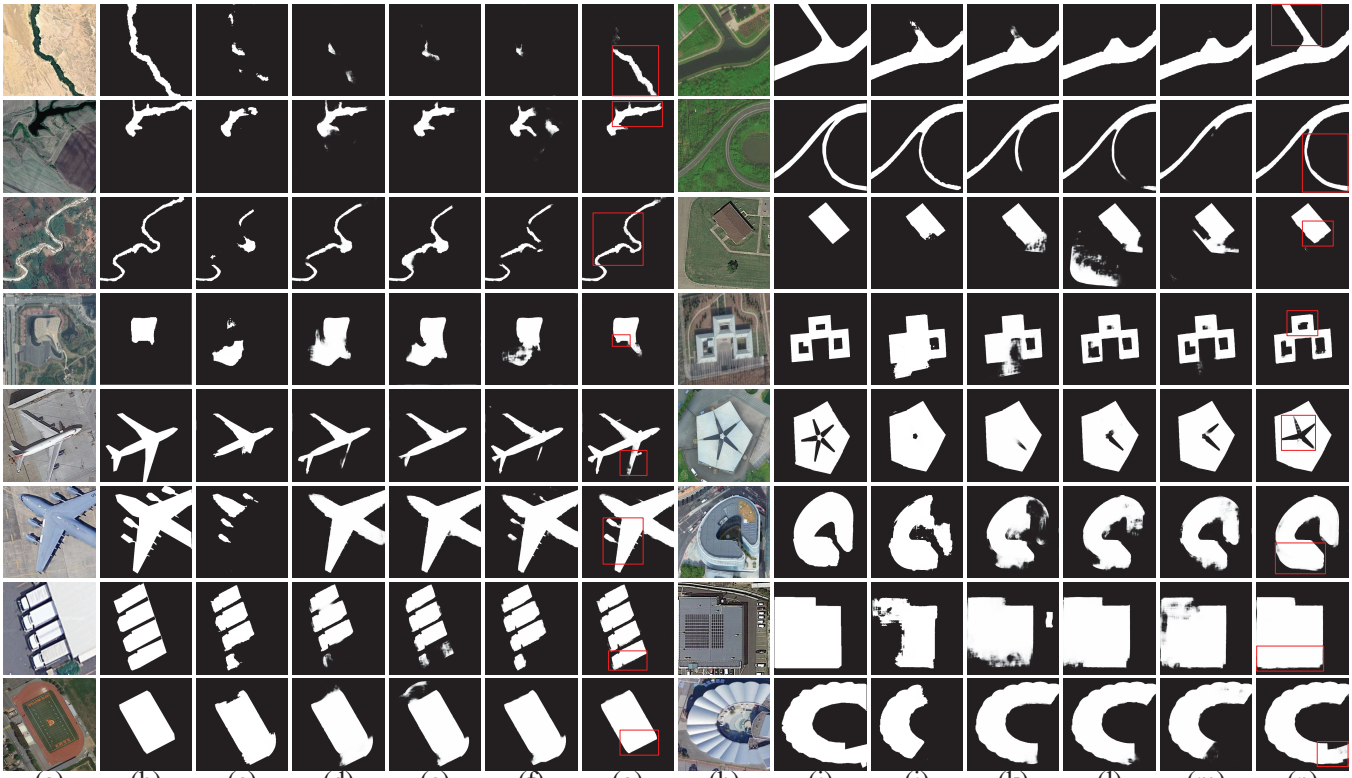


Fig. 10. Visual comparisons with various objective loss functions on the ORSI-4199 dataset. (a), (h) optical RSIs. (b), (i) GT. (c), (j) F-measure Loss. (d), (k) BCE Loss. (e), (l) CT Loss. (f), (m) ACT Loss. (g), (n) IGLoss.

TABLE VIII
EXPERIMENTAL RESULTS OF THE PRESENTED HFANET-R UTILIZING DIFFERENT LOSS FUNCTIONS ON THE ORSI-4199 DATASET.

Objective Loss Functions	MAE↓	F_β↑	S_m↑
F-measure Loss [37]	0.0510	0.7942	0.8211
BCE Loss	0.0395	0.8359	0.8639
CT Loss [38]	0.0379	0.8379	0.8647
ACT Loss [39]	0.0360	0.8469	0.8664
proposed IG Loss	**0.0343**	**0.8532**	**0.8709**
F-measure Loss [37] + IoU Loss	0.0442	0.8192	0.8366
BCE Loss + IoU Loss	0.0358	0.8525	0.8713
CT Loss [38] + IoU Loss	0.0352	0.8536	0.8710
ACT Loss [39] + IoU Loss	0.0345	0.8531	0.8696
proposed IG Loss + IoU Loss	**0.0329**	**0.8590**	**0.8758**

i.e., edge prediction and salient region prediction, thus lacking consistent mutual supervision. Based on this deficiency, we propose IGLoss that combines boundary detection and salient region detection via an adaptive weighting mechanism. As shown in Table VIII, when we utilize various objective functions to train the proposed HFANet-R, the gap in quantitative metrics is obvious. First, IGLoss without IoU Loss as auxiliary function can make the network converge to the best performance. It is worth noting that F-measure Loss does not calculate the cross-entropy of each pixel like BCE Loss, so the final performance is lower than that of BCE. And IGLoss applies the adaptive weights learned by the network to interactively supervise the two sub-tasks, and exerts the intrinsic relation between them, so the detection results are better than CT and ACT Loss. Second, when all functions are combined with IoU Loss, the superiority of IGLoss is further highlighted. The proposed HFANet achieves state-of-the-art performance on the ORSI-4199 dataset with MAE of 0.0329 precisely under the joint supervision of IGLoss and IoU Loss. Finally, we visualize the predicted saliency maps supervised by different loss functions, as shown in Fig. 10. For typical complex topological objects or regions, i.e. rivers, roads, airplanes, man-made buildings, the proposed IGLoss can better identify the most complete edges than other supervised functions.

V. CONCLUSION

We present a hybrid feature aligned network for salient object detection in optical RSIs. The proposed model is equipped with the hybrid encoder, AFAM, GF-ASPP and IGLoss. The hybrid encoder combines the advantages of both CNN and Transformer components, which can better extract local and global context in complex scenes. AFAM and GF-ASPP use alignment and context refinement strategies, respectively, to integrate multiscale features extracted by the hybrid encoder. We also propose an interactive guidance objective function, named IGLoss. It resamples the weights of boundary regions in an adaptive manner, and interactively guides the learning process of saliency maps and edge auxiliary maps. Experiments reveal that our algorithm outperforms 18 state-of-the-art methods both quantitatively and qualitatively.

REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] Y. Zhai and M. Shah, "Visual Attention Detection in Video Sequences Using Spatiotemporal Cues," in *Proc. ACM Int. Conf. Multimedia (MM'06)*, 2006, pp. 815–824.

[3] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 1597–1604.

[4] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Saliency Detection by Multiple-Instance Learning," *IEEE Trans. Cybern. (T-CYB)*, vol. 43, no. 2, pp. 660–672, 2013.

[5] Z. Xiong, Y. Yuan, and Q. Wang, "ASK: Adaptively Selecting Key Local Features for RGB-D Scene Recognition," *IEEE Trans. Image Process. (T-IP)*, vol. 30, pp. 2722–2733, 2021.

[6] Y. Yuan, Z. Xiong, and Q. Wang, "ACM: Adaptive Cross-Modal Graph Convolutional Neural Networks for RGB-D Scene Recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 33, Jul. 2019, pp. 9176–9184.

[7] Z. Xiong, Y. Yuan, N. Guo, and Q. Wang, "Variational Context-Deformable ConvNets for Indoor Scene Parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2020, pp. 3992–4002.

[8] Q. Wang, T. Han, Z. Qin, J. Gao, and X. Li, "Multitask Attention Network for Lane Detection and Fitting," *IEEE Trans. Neural Netw. Learn. Syst. (T-NNLS)*, vol. 33, no. 3, pp. 1066–1078, 2022.

[9] Y. Yuan, Z. Xiong, and Q. Wang, "VSSA-NET: Vertical Spatial Sequence Attention Network for Traffic Sign Detection," *IEEE Trans. Image Process. (T-IP)*, vol. 28, no. 7, pp. 3423–3434, 2019.

[10] Y. Yuan, Z. Xiong, and Q. Wang, "An Incremental Framework for Video-Based Traffic Sign Detection, Tracking, and Recognition," *IEEE Trans. Intell. Transp. Syst. (T-ITS)*, vol. 18, no. 7, pp. 1918–1929, 2017.

[11] Y. Yuan, Y. Lu, and Q. Wang, "Tracking as a whole: Multi-target tracking by modeling group behavior with sequential detection," *IEEE Trans. Intell. Transp. Syst. (T-ITS)*, vol. 18, no. 12, pp. 3339–3349, 2017.

[12] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, and S. Kwong, "Nested Network With Two-Stream Pyramid for Salient Object Detection in Optical Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens. (T-GRS)*, vol. 57, no. 11, pp. 9156–9166, 2019.

[13] Q. Zhang, R. Cong, C. Li, M.-M. Cheng, Y. Fang, X. Cao, Y. Zhao, and S. Kwong, "Dense Attention Fluid Network for Salient Object Detection in Optical Remote Sensing Images," *IEEE Trans. Image Process. (T-IP)*, vol. 30, pp. 1305–1317, 2021.

[14] Z. Tu, C. Wang, C. Li, M. Fan, H. Zhao, and B. Luo, "ORSI Salient Object Detection via Multiscale Joint Region and Boundary Model," *IEEE Trans. Geosci. Remote Sens. (T-GRS)*, vol. 60, pp. 1–13, 2022.

[15] C.-I. Chang, "An Effective Evaluation Tool for Hyperspectral Target Detection: 3D Receiver Operating Characteristic Curve Analysis," *IEEE Trans. Geosci. Remote Sens. (T-GRS)*, vol. 59, no. 6, pp. 5131-5153, 2021.

[16] Y. Liu, Q. Li, Y. Yuan, Q. Du, and Q. Wang, "ABNet: Adaptive Balanced Network for Multiscale Object Detection in Remote Sensing Imagery," *IEEE Trans. Geosci. Remote Sens. (T-GRS)*, vol. 60, pp. 1–14, 2022.

[17] Y. Liu, Q. Li, Y. Yuan, and Q. Wang, "Single-shot Balanced Detector for Geospatial Object Detection," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2022, pp. 2529–2533.

[18] Y. Yuan, D. Ma, and Q. Wang, "Hyperspectral Anomaly Detection by Graph Pixel Selection," *IEEE Trans. Cybern. (T-CYB)*, vol. 46, no. 12, pp. 3123–3134, 2016.

[19] Q. Wang, W. Huang, Z. Xiong, and X. Li, "Looking Closer at the Scene: Multiscale Representation Learning for Remote Sensing Image Scene Classification," *IEEE Trans. Neural Netw. Learn. Syst. (T-NNLS)*, vol. 33, no. 4, pp. 1414–1428, 2022.

[20] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word–Sentence Framework for Remote Sensing Image Captioning," *IEEE Trans. Geosci. Remote Sens. (T-GRS)*, vol. 59, no. 12, pp. 10 532–10 543, 2021.

[21] X. Zhang, W. Huang, Q. Wang, and X. Li, "SSR-NET: Spatial–Spectral Reconstruction Network for Hyperspectral and Multispectral Image Fusion," *IEEE Trans. Geosci. Remote Sens. (T-GRS)*, vol. 59, no. 7, pp. 5953–5965, 2021.

[22] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local Deep Features for Salient Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6593–6601.

[23] Q. Hou, M.-M. Cheng, X. Hu, A. Borji *et al.*, "Deeply Supervised Salient Object Detection with Short Connections," *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)*, vol. 41, no. 4, pp. 815–828, 2019.

[24] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse Attention for Salient Object Detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, July 2018.

[25] J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A Simple Pooling-Based Design for Real-Time Salient Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3912–3921.

[26] T. Zhao and X. Wu, "Pyramid Feature Attention Network for Saliency Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3080–3089.

[27] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-Scale Interactive Network for Salient Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 9410–9419.

[28] Z. Wu, L. Su, and Q. Huang, "Stacked Cross Refinement Network for Edge-Aware Salient Object Detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 7263–7272.

[29] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 35–51.

[30] J. Wei, S. Wang, and Q. Huang, "F^{3}Net: Fusion, Feedback and Focus for Salient Object Detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 34, 2020, pp. 12 321–12 328.

[31] M. Ma, C. Xia, and J. Li, "Pyramidal Feature Shrinking for Salient Object Detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 35, 2021, pp. 2311–2318.

[32] C. Li, R. Cong, C. Guo, H. Li, C. Zhang, F. Zheng, and Y. Zhao, "A parallel down-up fusion network for salient object detection in optical remote sensing images," *Neurocomputing*, vol. 415, pp. 411–420, 2020.

[33] G. Li, Z. Liu, W. Lin, and H. Ling, "Multi-content complementation network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens. (T-GRS)*, vol. 60, pp. 1–13, 2022.

[34] 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 Trans. Geosci. Remote Sens. (T-GRS)*, vol. 60, pp. 1–11, 2022.

[35] Z. Huang, H. Chen, B. Liu, and Z. Wang, "Semantic-Guided Attention Refinement Network for Salient Object Detection in Optical Remote Sensing Images," *Remote Sens.*, vol. 13, no. 11, 2021.

[36] X. Zhou, K. Shen, Z. Liu, C. Gong, J. Zhang, and C. Yan, "Edge-Aware Multiscale Feature Integration Network for Salient Object Detection in Optical Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens. (T-GRS)*, vol. 60, pp. 1–15, 2022.

[37] K. Zhao, S. Gao, W. Wang, and M.-M. Cheng, "Optimizing the F-Measure for Threshold-Free Salient Object Detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 8848–8856.

[38] Z. Chen, H. Zhou, J. Lai, L. Yang, and X. Xie, "Contour-Aware Loss: Boundary-Aware Learning for Salient Object Segmentation," *IEEE Trans. Image Process. (T-IP)*, vol. 30, pp. 431–443, 2021.

[39] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 9138–9147.

[40] Q. Wang, Y. Yuan, and P. Yan, "Visual Saliency by Selective Contrast," *IEEE Trans. Circuits Syst. Video Technol. (T-CSVT)*, vol. 23, no. 7, pp. 1150–1155, 2013.

[41] G. Zhu, Q. Wang, Y. Yuan, and P. Yan, "Learning Saliency by MRF and Differential Threshold," *IEEE Trans. Cybern. (T-CYB)*, vol. 43, no. 6, pp. 2032–2043, 2013.

[42] G. Zhu, Q. Wang, and Y. Yuan, "Tag-Saliency: Combining bottom-up and top-down information for saliency detection," *Comput. Vis. Image Underst.*, vol. 118, pp. 40–49, 2014.

[43] L. Wang, H. Lu, X. Ruan, and M. Yang, "Deep Networks for Saliency Detection via Local Estimation and Global Search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2015, pp. 3183–3192.

[44] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2019, pp. 1448–1457.

[45] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 733–740.

[46] D. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-Measure: A New Way to Evaluate Foreground Maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 4558–4567.

[47] D. Zhao, J. Wang, J. Shi, and Z. Jiang, "Sparsity-guided Saliency Detection for Remote Sensing Images," *J. Appl. Remote Sens.*, vol. 9, pp. 1–14, Sep. 2015.

[48] L. Zhang, S. Wang, and X. Li, "Salient region detection in remote sensing images based on color information content," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2015, pp. 1877–1880.

[49] L. Zhang, Y. Wang, and Y. Sun, "Salient Target Detection Based on the Combination of Super-Pixel and Statistical Saliency Feature Analysis for Remote Sensing Images," in *Proc. IEEE Int. Conf. Inf. Process. (ICIP)*, 2018, pp. 2336–2340.
[50] L. Zhang, Y. Liu, and J. Zhang, "Saliency detection based on self-adaptive multiple feature fusion for remote sensing images," *Int. J. Remote Sens.*, vol. 40, no. 22, pp. 8270–8297, 2019.
[51] Z. Huang, H. Chen, T. Zhou, Y. Yang, C. Wang, and B. Liu, "Contrast-weighted dictionary learning based saliency detection for VHR optical remote sensing images," *Pattern Recognit.*, vol. 113, p. 107757, 2021.
[52] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–15.
[53] W. Wang, E. Xie, X. Li, P. Fan *et al.*, "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, October 2021, pp. 568–578.
[54] W. Wang, E. Xie, X. Li, P. Fan, *et al.*, "PVT v2: Improved baselines with Pyramid Vision Transformer," *Computational Visual Media*, 2022.
[55] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," *arXiv preprint arXiv:2102.04306*, 2021.
[56] Y. Gao, M. Zhou, and D. Metaxas, "UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention (MICCAI)*, 2021, pp. 61–71.
[57] Y. Sha, Y. Zhang, X. Ji, and L. Hu, "Transformer-Unet: Raw Image Processing with Unet," *arXiv preprint arXiv:2109.08417*, 2021.
[58] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-wise Perspective with Transformer," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2022.
[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017, pp. 1–11.
[60] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)," *arXiv preprint arXiv:1606.08415*, 2016.
[61] Z. Huang, Y. Wei, X. Wang, W. Liu, T. S. Huang, and H. Shi, "AlignSeg: Feature-Aligned Segmentation Networks," *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)*, vol. 44, no. 1, pp. 550–557, 2022.
[62] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable Convolutional Networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct 2017, pp. 764–773.
[63] Y. Chen, C. Han, N. Wang, and Z. Zhang, "Revisiting feature alignment for one-stage object detection," *arXiv preprint arXiv:1908.01570*, 2019.
[64] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Guided Upsampling Network for Real-Time Semantic Segmentation," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, September 2018.
[65] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial Transformer Networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 28, 2015, pp. 1–9.
[66] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)*, vol. 40, no. 4, pp. 834–848, 2018.
[67] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2999–3007.
[68] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process. (T-IP)*, vol. 13, no. 4, pp. 600–612, 2004.
[69] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6230–6239.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and remote sensing.

Yanfeng Liu (Student Member, IEEE) received the B.E. degree in computer science and technology from Northeast Forestry University, Harbin, China, in 2021. He is pursuing the M.S. degree in computer science and technology with the School of Computer Science and School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China.

His current research interests include computer vision, pattern recognition and remote sensing.

Zhitong Xiong (Member, IEEE) received the Ph.D. degree in computer science and technology from the Northwestern Polytechnical University, Xi'an, China, in 2021. He is currently a Postdoc with the Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany.

His research interests include computer vision, machine learning and remote sensing.

Yuan Yuan (M'05-SM'09) is currently a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS AND PATTERN RECOGNITION, as well as the conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.