

Spectrum-driven Mixed-frequency Network for Hyperspectral Salient Object Detection

Peifu Liu, Tingfa Xu[†], Huan Chen, Shiyun Zhou, Haolin Qin, Jianan Li[†]

Abstract—Hyperspectral salient object detection (HSOD) aims to detect spectrally salient objects in hyperspectral images (HSIs). However, existing methods inadequately utilize spectral information by either converting HSIs into false-color images or converging neural networks with clustering. We propose a novel approach that fully leverages the spectral characteristics by extracting two distinct frequency components from the spectrum: low-frequency Spectral Saliency and high-frequency Spectral Edge. The Spectral Saliency approximates the region of salient objects, while the Spectral Edge captures edge information of salient objects. These two complementary components, crucial for HSOD, are derived by computing from the inter-layer spectral angular distance of the Gaussian pyramid and the intra-neighborhood spectral angular gradients, respectively. To effectively utilize this dual-frequency information, we introduce a novel lightweight Spectrum-driven Mixed-frequency Network (SMN). SMN incorporates two parameter-free plug-and-play operators, namely Spectral Saliency Generator and Spectral Edge Operator, to extract the Spectral Saliency and Spectral Edge components from the input HSI independently. Subsequently, the Mixed-frequency Attention module, comprised of two frequency-dependent heads, intelligently combines the embedded features of edge and saliency information, resulting in a mixed-frequency feature representation. Furthermore, a saliency-edge-aware decoder progressively scales up the mixed-frequency feature while preserving rich detail and saliency information for accurate salient object prediction. Extensive experiments conducted on the HS-SOD benchmark and our custom dataset HSOD-BIT demonstrate that our SMN outperforms state-of-the-art methods regarding HSOD performance. Code and dataset will be available at <https://github.com/laprf/SMN>.

Index Terms—Hyperspectral salient object detection, Spectrum, Mixed-frequency Attention

I. INTRODUCTION

HYPERSPECTRAL imaging systems offer a unique capability to capture data from observed scenes, providing both high spatial resolution and abundant spectral information. This enables the acquisition of hyperspectral images (HSIs) consisting of numerous contiguous narrow spectral bands [1]. By selecting an arbitrary point from the spatial dimension of the hyperspectral cube, the spectral response curve effectively represents the distinctive characteristics of a target. This characteristic has led to the increasing significance of HSIs in various disciplines, including target detection [2], spectral estimation [3], and remote sensing [4].

Peifu Liu, Tingfa Xu, Huan Chen, Shiyun Zhou, Haolin Qin, and Jianan Li are with Beijing Institute of Technology, Beijing 10081, China. Email: {laprf, ciom_xtf1, huanchen, zhoushiyun, 3120225333, lijianan}@bit.edu.cn

Tingfa Xu is also with the Key Laboratory of Photoelectronic Imaging Technology and System, Ministry of Education of China, Beijing 100081, China, and with the Chongqing Innovation Center, Chongqing 401135, China.

[†] Correspondence to: Tingfa Xu and Jianan Li.

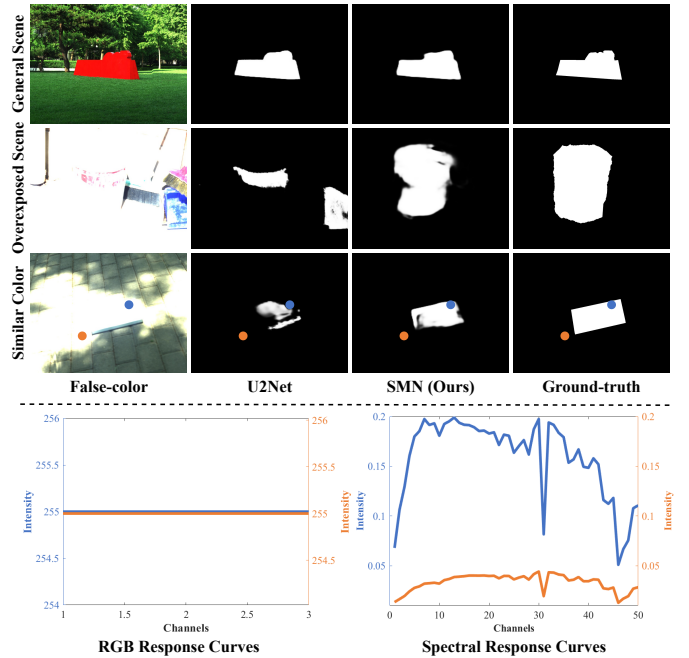


Fig. 1. Comparison with an RGB-based saliency detection method, U2Net [5]. The input of U2Net is false-color images. We compare the RGB and spectral response curves of the foreground point (blue) and background point (orange). The RGB response remains consistent across all three channels, leading to detection failure. However, the spectral curves of these two points exhibit distinct characteristics, enabling their discrimination using spectrum information. Consequently, our proposed Spectrum-driven Mixed-frequency Network (SMN) outperforms U2Net, particularly in the latter two scenarios.

By simulating human visual attention, salient object detection (SOD) aims to locate and segment the most salient object or region in a scene [6]. Recent advancements in SOD have witnessed the utilization of convolutional neural networks [7], [8] or Transformers [9], [10], which have significantly enhanced the representation capability of features, resulting in notable improvements in detection performance [11]. However, traditional SOD algorithms heavily rely on color information to discriminate foreground and background objects. In certain exceptional cases, such as scenes with overexposure or situations where the foreground and background colors are similar, the color information of the foreground object may be absent or indistinguishable from the background, leading to detection failures (as illustrated in Figure 1).

In contrast to RGB images, HSIs offer a wealth of spectral information, enabling a more comprehensive characterization of an object's material, composition, and other intrinsic properties. HSIs exhibit a higher resilience to variations in

illumination conditions and are less reliant on color and texture information typically present in RGB images. For instance, as depicted in Figure 1, we have selected a foreground point (blue) and a background point (orange) and plotted their respective spectral and RGB response curves. The bottom section of Figure 1 clearly illustrates that the RGB response remains consistent across all three channels, making it challenging for false-color images to differentiate between the two points effectively. However, the spectral curves of these two points exhibit distinct characteristics, enabling their discrimination using spectrum information. Hyperspectral salient object detection (HSOD), which involves detecting the most salient object within HSIs, therefore holds tremendous potential for diverse applications, including pest control [12], military surveillance [13], and environmental management [14].

Traditional methods [15]–[17] used for HSOD often rely on a simple conversion of HSIs into false-color images, which only exploit a fraction of the available spectral data. Recently, some deep learning techniques [1], [18] directly employ convolutional neural networks (CNNs) to extract features of entire HSIs and subsequently utilize clustering algorithms to generate saliency maps. Although these methods have achieved remarkable results, they usually encounter several limitations: i) Clustering algorithms are often time-consuming; ii) CNNs extract a large number of features, not all of which are useful for saliency detection, leading to computational redundancy and further burdening time resources; iii) The applicability of clustering algorithms on CNN-extracted features is limited, resulting in suboptimal detection performance compared to end-to-end networks.

To address the above issues, we conducted an extensive investigation of HSIs and identified two distinct frequency components that can be extracted based on the spectrum, playing a crucial role in the accurate identification and localization of salient objects. Specifically, the lower-frequency component, known as Spectral Saliency, provides an approximate localization of the salient target, while the higher-frequency component, known as Spectral Edge, enhances the edges of the target. Motivated by the aim of fully leveraging these two types of spectrum-extracted information that complement each other, we introduce the Spectrum-driven Mixed-frequency Network (SMN), the first end-to-end deep network designed for HSOD. SMN is a lightweight model, which comprises four key components: the extraction of Spectral Saliency and Spectral Edge maps, frequency-specific embeddings, a Mixed-frequency Attention module, and a decoder aware of both saliency and edge information.

Firstly, we extract the Spectral Saliency and Spectral Edge maps from the input HSIs using two plug-and-play operators: the Spectral Saliency Generator (SSG) and the Spectral Edge Operator (SEO). In the SSG, we construct a spatially blurred Gaussian pyramid, with each layer containing the complete spectrum of the input. The spectral angular distance (SAD) between the pyramid layers is utilized to compute Spectral Saliency maps. In the SEO, we first calculate the SAD between each pixel and its neighborhood. Subsequently, we employ various kernels to compute the gradient of the SAD

values within the neighborhood, enabling the determination of the Spectral Edge maps. It is worth noting that these two operators perform their computations rapidly without the need for learnable parameters.

Secondly, Spectral Saliency and Spectral Edge images are subsequently transformed into deep features employing frequency-specified embeddings. By leveraging a low-frequency embedding process, deep saliency features are obtained. For the high-frequency embedding, we employ cascaded convolutional layers along with an Edge Detection Module, which combines to capture intricate edge details. Both low-frequency and high-frequency embeddings generate deep features related to saliency and edge, respectively. Considering that edge features are considered low-level features, the high-frequency embedding is designed to be shallower than the low-frequency embedding. This not only enhances the efficiency of the feature extraction process but also minimizes the number of required parameters.

Thirdly, we introduce a Mixed-Frequency Attention module designed to enable intricate interactions and fusion among multi-frequency deep features, resulting in a nuanced mixed-frequency feature representation. The module comprises two frequency-dependent heads: a low-frequency head and a high-frequency head. The low-frequency head ingests saliency features and employs self-attention mechanisms to accentuate regions of importance. Conversely, the high-frequency head processes both saliency and edge features, executing cross-attention between them. Given that the low-frequency saliency representation is refined in light of neighboring high-frequency edge details, and that unrestricted long-range interactions between frequencies could introduce noise or be counterproductive, we opt for a localized attention paradigm—specifically, the neighborhood attention mechanism—as elaborated in NAT [19]. This approach not only ensures contextually relevant interactions within a confined neighborhood but also mitigates computational burden by maintaining linear complexity, unlike the quadratic complexity inherent in traditional attention mechanisms. The synthesized mixed-frequency feature encapsulates a rich amalgamation of both edge and saliency attributes.

Finally, we employ a saliency-edge-aware decoder that progressively upscales the mixed-frequency feature. Since the shallow saliency information and edge details from the frequency-specified embedding phase are simultaneously preserved, a saliency map exhibits high-fidelity edges and superior detection accuracy can be obtained. Consequently, the resulting saliency maps exhibit superior detection performance while minimizing computational costs.

We conducted extensive evaluations of our model on the HS-SOD dataset [20] as well as our collected dataset, HSOD-BIT. The results demonstrate that our proposed method surpasses the existing state-of-the-art HSOD methods. Our model is lightweight yet capable of detecting salient objects more comprehensively compared to RGB-based SOD methods, particularly in scenarios with overexposure and similar foreground and background colors.

Our contributions can be summarized as follows:

- We propose a novel Spectrum-driven Mixed-frequency Network for the task of HSOD. This approach effectively

utilizes both low-frequency and high-frequency information present in HSIs to detect salient objects. To our knowledge, our work represents the first attempt to apply an end-to-end neural network to the HSOD problem.

- We introduce two parameter-free plug-and-play operators, namely the Spectral Saliency Generator and the Spectral Edge Operator, specifically designed for HSIs. These operators enable us to leverage spectral information and provide frequency-specific information effectively.
- We tailor a Mixed-frequency Attention module to fully exploit the distinct frequency properties present in HSIs. The design of frequency-dependent heads enables the network to focus on different types of information.
- We present quantitative and qualitative experimental results that demonstrate the superiority of our method compared to state-of-the-art HSOD methods on both the HS-SOD and HSOD-BIT datasets.

II. RELATED WORK

A. Salient Object Detection

Conventional SOD methods rely on hand-crafted features [15], [21], [22]. For instance, Rosin *et al.* [21] employ edge detection, threshold decomposition, and pixel-wise operations to identify salient objects. Alexe *et al.* [22] utilize a generic objectness prior by leveraging object proposals. While these hand-crafted features enable real-time SOD, they have several limitations in capturing salient objects in complex scenarios [23]. For instance, they tend to emphasize high-contrast edges rather than the salient object itself, and the preservation of boundaries is often inadequate [24].

In contrast, CNNs possess exceptional feature extraction capabilities and can identify the most salient regions without relying on prior knowledge [7], [8]. Li *et al.* [7] separate the encoding of low-level and high-level information, flatten and concatenate them, and then input the resulting data into a two-layer perceptron to predict the saliency region. Zhang *et al.* [8] employ saliency cues and a multi-level fusion mechanism to detect salient objects. Yao *et al.* [25] integrate the edge extraction module with the prediction network, yielding saliency maps with precise edge delineation. However, these SOD techniques are limited to RGB data and cannot be directly applied to hyperspectral data for HSOD.

B. Hyperspectral Salient Object Detection

Despite the significant advancements in SOD, the field of hyperspectral imaging remains relatively new in this context. Wilson *et al.* [26] introduced the concept of contrast sensitivity saliency to fuse different bands and visualize hyperspectral remote sensing images. Subsequently, dimension reduction techniques and Itti's attention model [15] were incorporated into HSOD. Moan *et al.* [17] divided the spectrum of a hyperspectral image into three regions and employed principal component analysis (PCA) to extract the first principal component of each region. Zhang *et al.* [27] utilized both real color and PCA images for visualization. However, although these methods offer computational efficiency, they suffer from inevitable information loss due to feature reduction.

Recently, Imamoglu *et al.* [18], [20] introduced the first dataset specifically designed for HSOD. They employed a manifold ranking algorithm and extracted features using a self-supervised CNN to generate saliency maps. Similarly, Huang *et al.* [1] utilized a CNN with two channels to extract spatial and spectral features separately, which were subsequently fused to optimize the saliency values of both foreground and background cues, leading to improved detection performance but at the expense of high computational complexity as well as low computing speed. Our model takes Spectral Saliency and Spectral Edge as high-level inputs. Instead of employing clustering algorithms, it employs an end-to-end neural network for salient object detection, effectively addressing the above-mentioned drawbacks.

C. Attention Mechanism in Salient Object Detection

Although the attention mechanism is crucial for SOD, it was first employed in image classification [28]. Yin *et al.* [29] were the first attempt to incorporate the attention mechanism into CNNs. The subsequent emergence of the self-attention mechanism demonstrated a powerful ability to capture features with long-range dependencies, witnessing great success in machine translation [30] and image classification [31], *etc.*

The potential of the self-attention mechanism in SOD was first recognized by Liu *et al.* [9] and Zhang *et al.* [10]. Subsequent studies [32]–[34] have further expanded the application of self-attention in SOD tasks. In our work, we introduce a novel attention mechanism called Mixed-frequency Attention, which employs one attention head to concentrate solely on saliency information while another focuses on the interaction between edge and saliency information. This pioneering approach represents the first integration of the attention mechanism into the HSOD task.

III. METHOD

Given an HSI denoted as $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, the primary objective of hyperspectral salient object detection is to generate a saliency map denoted as $\mathbf{S} \in \mathbb{R}^{H \times W \times 1}$, which provides information about the location of the salient object within the HSI. Such a mapping process can be formulated as:

$$\mathbf{S} = \Phi(\mathbf{I}). \quad (1)$$

The mapping function $\Phi(\cdot)$ is implemented by a novel Spectrum-driven Mixed-frequency Network (SMN).

Figure 2 (a) presents the comprehensive architecture of SMN, which encompasses four key steps: Spectral Saliency and Spectral Edge extraction, frequency-specified embeddings, Mixed-frequency Attention, and saliency-edge-aware decoding. To elaborate, the first step involves the extraction of Spectral Saliency and Spectral Edge images using dedicated plug-and-play operators. These images are subsequently incorporated into deep saliency or edge features by means of frequency-specified embeddings. Moving on to the third step, a Mixed-frequency Attention module is employed to fully harness the complementary nature of these features. This enables the generation of a mixed-frequency feature that encompasses comprehensive edge and saliency information. Lastly,

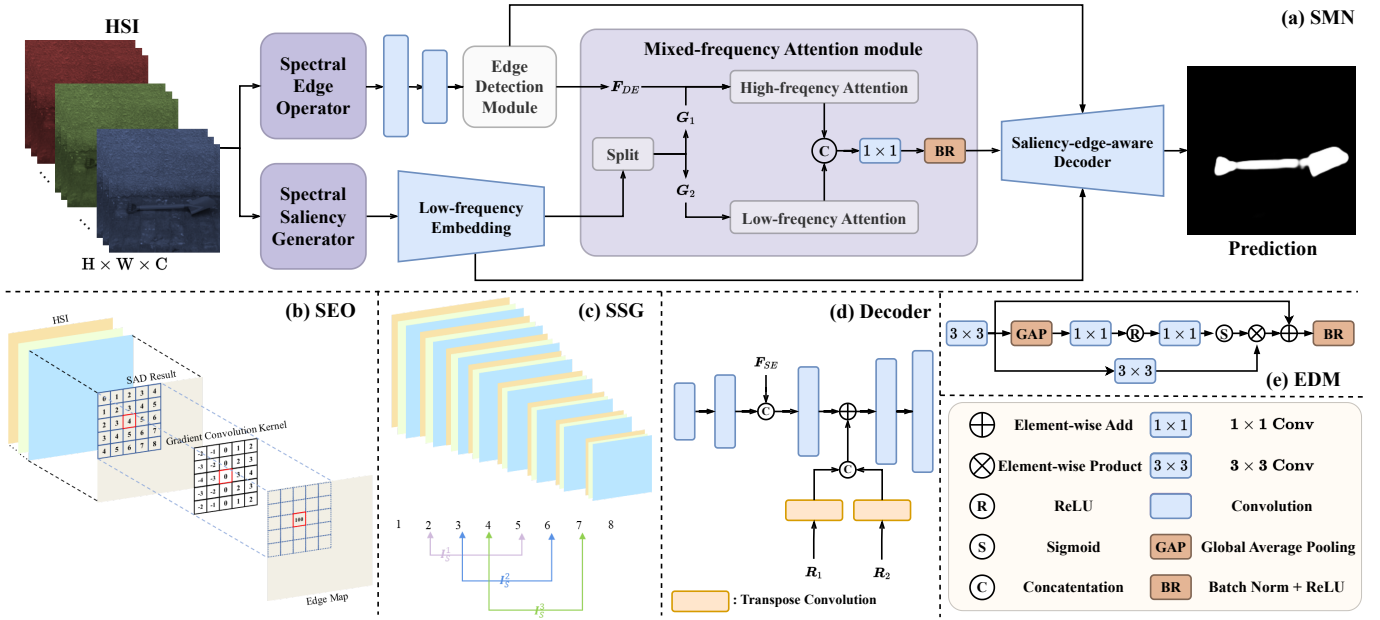


Fig. 2. (a) Illustration of the Spectrum-Driven Mixed-Frequency Network (SMN) employing an encoder-bottleneck-decoder architecture. The encoder comprises two distinct modules: the Spectral Edge Operator (SEO) and the Spectral Saliency Generator (SSG). The bottleneck integrates a Mixed-Frequency Attention Module, featuring frequency-dependent attention heads. The deep edge feature is denoted by F_{DE} , whereas G_1 and G_2 denote split saliency features. (b) SEO detects edge information by calculating the spectral angular distance (SAD) result’s gradient. (c) SSG generates saliency maps I_S^1 , I_S^2 , and I_S^3 by estimating the difference between pyramid levels. (d) The decoder preserves low-level encoder features R_1 and R_2 , and the shallow edge feature F_{SE} to generate better saliency maps. (e) Edge Detection Module (EDM) generates an edge feature.

in the fourth step, the saliency-edge-aware decoder gradually upscales the mixed-frequency feature while simultaneously preserving the fine-grained edge details and shallow saliency information. Ultimately, this decoding process culminates in the production of the final saliency map.

A. Spectral Saliency Generator

The Spectral Saliency Generator (SSG) is a stand-alone layer responsible for generating Spectral Saliency maps. As illustrated in Figure 2 (c), these maps are produced by computing the “center-surround” similarity between pairs of Gaussian pyramid layers, constructed from the input HSI. The Spectral Saliency maps provide an approximate indication of the salient object’s location and serve as the low-frequency input to SMN.

Specifically, the input HSI undergoes an initial downsampling process using Gaussian downsampling operations. This process involves applying depth-wise convolution with a fixed Gaussian weight to create a Gaussian pyramid with N layers ($N = 8$). Through Gaussian downsampling, the spatial dimensions of the image decrease as the scale increases, and each pixel’s information is influenced by a larger neighborhood of pixels. This enables the assessment of the saliency value between a “center” pixel at point (i, j) and its “surround” pixel. The comparison is executed via the calculation of the spectral angular distance (SAD) between spectral vectors \mathbf{v}_c and \mathbf{v}_s , which are derived from the c -th and s -th layers of the Gaussian pyramid, respectively. The value of the saliency map I_S at this point is computed as follows:

$$I_S(i, j) = \arccos \left(\frac{\mathbf{v}_c^T(i, j) \mathbf{v}_s(i, j)}{\|\mathbf{v}_c(i, j)\| \|\mathbf{v}_s(i, j)\|} \right), \quad (2)$$

where $\|\cdot\|$ is the Euclidean norm of a vector. In this context, the layer index c of the “center” pixel takes on values from the set $\{2, 3, 4\}$, and s is determined as $c + 3$. By performing the aforementioned calculation for each point of the image, the saliency map of the entire HSI can be obtained. Three values of the layer index c yield different saliency maps, denoted as $\{I_S^k\}_{k=1}^3$. The dimensions of each saliency map are $H \times W \times 1$.

B. Spectral Edge Operator

The blurring of object edges and loss of high-frequency information in Spectral Saliency images are consequences of Gaussian downsampling. Employing such images for saliency detection could result in less sharp or even erroneous edges in the detection results. To incorporate high-frequency details into the SMN, we have devised a module called the Spectral Edge Operator (SEO). Drawing inspiration from edge detection operators like the Canny operator, SEO extracts Spectral Edge images by computing the gradient of the SAD in the vicinity of each pixel.

Specifically, for a point (i, j) on the HSI, assume its neighborhood size is $H' \times W'$. The SAD between this point and any point (p, q) within its neighborhood can be computed, resulting in a local spectral similarity map $M \in \mathbb{R}^{H' \times W'}$. This process can be formulated as follows:

$$M(p, q) = \arccos \left(\frac{\mathbf{v}(i, j)^T \mathbf{v}(p, q)}{\|\mathbf{v}(i, j)\| \|\mathbf{v}(p, q)\|} \right), \quad (3)$$

where $\mathbf{v}(i, j)$ and $\mathbf{v}(p, q)$ represent the spectral vectors of points at (i, j) and (p, q) , respectively. To compute the value of the edge image I_E at (i, j) , gradient convolution kernels

the same size as the neighborhood, denoted as G_x and G_y , are applied to the local spectral similarity map M :

$$I_E(i, j) = |G_x * M| + |G_y * M|. \quad (4)$$

In Figure 2 (b), the specific content of G_x is depicted, with a size of 5. G_y is the transpose of G_x . By applying these operations to every pixel in the image, the Spectral Edge image for the entire image can be obtained. Efficient CUDA kernels of varying sizes are employed to expedite the computation. Three kernels of varying sizes are employed to extract Spectral Edge images $\{I_E^k\}_{k=1}^3$, where the dimensions of each image are $H \times W \times 1$.

C. Frequency-specified Embeddings

Both the Spectral Saliency Generator and the Spectral Edge Operator produce three sets of Spectral Saliency and Spectral Edge maps, respectively. These maps capture distinct and important information for hyperspectral salient object detection. To leverage this valuable information, the maps are transformed into deep saliency or edge features using frequency-specified embeddings.

Deep Saliency Feature. The deep saliency feature is obtained through a low-frequency embedding process. To generate this feature, the Spectral Saliency images, denoted as I_S^1 , I_S^2 , and I_S^3 , are concatenated along the channel dimension:

$$F_S = [I_S^1, I_S^2, I_S^3], \quad (5)$$

where F_S is the concatenation result. By concatenating the Spectral Saliency images, rather than sequentially feeding them into the network, the computational complexity is reduced, leading to improved inference speed. F_S are then transformed into a deep saliency feature F_{DS} as:

$$F_{DS} = f_L(F_S), \quad (6)$$

where $f_L(\cdot)$ represents the low-frequency embedding process.

Let $R = \{R_i | i = 1, 2, 3\}$ represent the multi-stage features obtained from the low-frequency embedding process, where each stage captures $\frac{1}{4}$, $\frac{1}{8}$, and $\frac{1}{16}$ of the input feature, respectively. The deep saliency feature refers to the last stage of the low-frequency embedding: $F_{DS} = R_3 \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C_S}$, where C_S is the channel dimension of the deep saliency feature. Furthermore, the shallow saliency information captured in R_1 and R_2 is retained for later use in the decoder stage.

Deep Edge Feature. A deep edge feature can be obtained through high-frequency embedding, which incorporates a downsampling block and an Edge Detection Module (EDM) [35]. The high-frequency embedding takes concatenated Spectral Edge images as input and generates a shallow edge feature F_{SE} , which can be expressed as follows:

$$F_{SE} = f_D([I_E^1, I_E^2, I_E^3]), \quad (7)$$

where $f_D(\cdot)$ denotes the downsampling block implemented using two $conv3 \times 3$ layers with a stride of 2, followed by batch normalization layers and ReLU activation functions. The shallow edge feature F_{SE} is then transformed into a deep edge feature F_{DE} using EDM $f_E(\cdot)$:

$$F_{DE} = f_E(F_{SE}). \quad (8)$$

Here, the shape of F_{DE} is $\frac{H}{16} \times \frac{W}{16} \times C_E$, with C_E denoting the dimension of the edge feature. The specific composition of EDM is illustrated in Figure 2 (e).

To enhance the quality of the generated edge features, we transform F_{DE} into an edge image M_E , which is constrained by a ground truth edge image. The generation process of M_E can be mathematically described as follows:

$$M_E = f_{CU}(F_{DE}), \quad (9)$$

where $f_{CU}(\cdot)$ is implemented using a $conv1 \times 1$ layer followed by an upsampling layer. The resulting edge map M_E is a single-channel grayscale image with the same dimensions as the original image. The ground truth edge image E is generated using an edge detector [36], which is obtained by combining two edge maps:

$$E = e(I_{FC}) + e(I'_S), \quad (10)$$

where $e(\cdot)$ represents the edge detector. I_{FC} corresponds to the false-color image rendered from HSI, while I'_S is the sum of the previously generated spectral saliency maps I_S .

D. Mixed-frequency Attention

The Mixed-frequency Attention (MA) module facilitates the comprehensive interaction and fusion of deep features with different frequencies. It encompasses two essential components: the high-frequency head and the low-frequency head. In the high-frequency head, a cross-attention mechanism is employed to capture the relationship between high-frequency edge features and low-frequency saliency features. This interaction enables the refinement of saliency representations under the constraints imposed by edge information. Conversely, in the low-frequency head, a self-attention mechanism is applied to the saliency feature itself. This enables the generation of more precise and accurate saliency representations by emphasizing relevant saliency information within the feature.

Saliency Feature Division. The saliency feature, denoted by F_{DS} and having a channel dimension C_S , is uniformly partitioned into two groups along the channel dimension:

$$\begin{aligned} G_1 &= F_{DS} \left(:, :, 0 : \lfloor \frac{C_S}{2} \rfloor \right), \\ G_2 &= F_{DS} \left(:, :, \lfloor \frac{C_S}{2} \rfloor : C_S \right), \end{aligned} \quad (11)$$

where G_1 and G_2 represent the resulting groups after the division. Each group has dimensions of $\frac{H}{16} \times \frac{W}{16} \times \lfloor \frac{C_S}{2} \rfloor$. These groups are subsequently fed into different heads of the Mixed-frequency Attention module for further processing.

High-frequency Attention Head. The high-frequency attention head incorporates both the deep saliency and edge features, enabling their interaction to enhance the accuracy of saliency detection. We employ a neighborhood attention mechanism (NAM) [19] to confine the receptive field of the *query* to its local neighborhood, enhancing its sensitivity to edge information while reducing computational complexity. Suppose the input matrices of the NAM are denoted as X and Y , respectively. The NAM can be defined as:

$$f_{NAM}(X, Y) = \sigma \left(Q_{i,j} K_{\rho(i,j)}^T + B_{i,j} \right) V_{\rho(i,j)}, \quad (12)$$

where $\rho(i, j)$ represents the neighborhood of a pixel at position (i, j) , and $B_{i,j}$ denotes the relative positional bias. The function $\sigma(\cdot)$ corresponds to the Sigmoid function. The *query* matrix \mathbf{Q} is derived from \mathbf{X} , while the *key* matrix \mathbf{K} and the *value* matrix \mathbf{V} are obtained from \mathbf{Y} :

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^Q, \mathbf{K} = \mathbf{Y}\mathbf{W}^K, \mathbf{V} = \mathbf{Y}\mathbf{W}^V, \quad (13)$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ represent learnable parameters implemented through linear projection.

The cross-attention between the deep edge feature \mathbf{F}_{DE} and the first group of deep saliency feature \mathbf{G}_1 is computed:

$$\mathbf{F}_H = \mathbf{f}_{\text{NAM}}(\mathbf{F}_{DE}, \mathbf{G}_1), \quad (14)$$

where \mathbf{F}_H represents the refined saliency feature, constrained by edge information.

Low-frequency Attention Head. The low-frequency attention heads exclusively receive deep saliency features, utilizing a self-attention mechanism to capture more precise representations of salient objects. To alleviate computational complexity, the NAM is employed. The self-attention operation is employed for the low-frequency attention result \mathbf{F}_L :

$$\mathbf{F}_L = \mathbf{f}_{\text{NAM}}(\mathbf{G}_2, \mathbf{G}_2). \quad (15)$$

After the low-frequency attention, \mathbf{F}_L serves as a more accurate saliency representation in comparison to the input \mathbf{G}_2 . It should be noted that the kernel size of the NAM differs between the two heads, reflecting the differences in input and objectives for the different frequency heads.

Frequency Convergence. To integrate the frequency-specific features, the outputs from the high-frequency and low-frequency attention heads are concatenated along the feature dimension, obtaining a mixed-frequency feature \mathbf{F}_{out} that combines comprehensive and effectively integrated edge and saliency information:

$$\mathbf{F}_{\text{out}} = \delta(\mathbf{f}_C([\mathbf{F}_H, \mathbf{F}_L])), \quad (16)$$

where $\delta(\cdot)$ is implemented by the ReLU activation function, and \mathbf{F}_C represents the $\text{conv}1 \times 1$ operation.

E. Saliency-edge-aware Decoder

As illustrated in Figure 2 (d), the saliency-edge-aware decoder employs a cascading structure of convolutional layers, allowing for the gradual upscaling of the mixed-frequency feature. This process ensures the fusion of shallow features from the encoder, preserving intricate details and saliency information.

The cascaded decoder architecture consists of five convolutional layers, each accompanied by a batch normalization layer, a ReLU activation function, and an interpolation operation. Let $\mathbf{D}_{\text{in}} \in \{\mathbf{D}_{\text{in}}^i | i = 1, 2, 3, 4, 5\}$ and $\mathbf{D}_{\text{out}} \in \{\mathbf{D}_{\text{out}}^i | i = 1, 2, 3, 4, 5\}$ represent the input and output of these convolutional layers, respectively. To preserve shallow information, the shallow edge information \mathbf{F}_{SE} is concatenated with $\mathbf{D}_{\text{out}}^2$ along the channel dimension:

$$\mathbf{D}_{\text{in}}^3 = [\mathbf{D}_{\text{out}}^2, \mathbf{F}_{SE}]. \quad (17)$$

The saliency information \mathbf{R}_1 and \mathbf{R}_2 are separately upsampled to match the spatial dimension, and the resulting outputs are concatenated in the channel dimension and added to $\mathbf{D}_{\text{out}}^3$:

$$\mathbf{D}_{\text{in}}^4 = \mathbf{D}_{\text{out}}^3 + [\mathbf{f}_{\text{tc1}}(\mathbf{R}_1), \mathbf{f}_{\text{tc2}}(\mathbf{R}_2)], \quad (18)$$

where $\mathbf{f}_{\text{tc1}}(\cdot)$ and $\mathbf{f}_{\text{tc2}}(\cdot)$ represent transposed convolutional layers. By incorporating edge information through concatenation and saliency information through summation, we reduce the modification of shallow information, resulting in a less complex network that retains more shallow information.

F. Hybrid Loss Function

During the training process, certain intermediate results are supervised to ensure the precise extraction of saliency or edge features. To achieve this, a hybrid loss function \mathcal{L} is utilized, given by the equation:

$$\mathcal{L} = \mathcal{L}_{\text{edge}} + \mathcal{L}_{\text{final}}, \quad (19)$$

where $\mathcal{L}_{\text{edge}}$ and $\mathcal{L}_{\text{final}}$ represent the loss associated with edge detection and the final saliency map, respectively. Further details will be provided subsequently.

Binary Cross-entropy Loss. The binary cross-entropy (BCE) loss function is defined as:

$$\mathcal{L}_{\text{BCE}}(\mathbf{X}, \mathbf{Y}) = - \sum [\mathbf{X} \log(\mathbf{Y}) + (1 - \mathbf{X}) \log(1 - \mathbf{Y})], \quad (20)$$

where \mathbf{X} represents the ground-truth values and \mathbf{Y} corresponds to the input matrix. In the case of edge map \mathbf{M}_E , it is supervised using the BCE loss with the edge ground truth \mathbf{E} as follows:

$$\mathcal{L}_{\text{edge}} = \mathcal{L}_{\text{BCE}}(\mathbf{M}_E, \mathbf{E}). \quad (21)$$

Intersection Over Union Loss. In accordance with Qin *et al.* [37], we incorporate the intersection over union (IoU) loss function, defined as:

$$\mathcal{L}_{\text{IoU}}(\mathbf{X}, \mathbf{Y}) = 1 - \frac{\sum_{r=1}^H \sum_{c=1}^W \mathbf{X}(r, c) \mathbf{Y}(r, c)}{\sum_{r=1}^H \sum_{c=1}^W [\mathbf{X}(r, c) + \mathbf{Y}(r, c) - \mathbf{X}(r, c) \mathbf{Y}(r, c)]}, \quad (22)$$

where \mathbf{X} and \mathbf{Y} denote the input and ground-truth matrices, respectively, and H and W represent the height and width. This loss function is employed to supervise the final saliency map \mathbf{S} :

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{IoU}}(\mathbf{S}, \mathbf{G}) + \mathcal{L}_{\text{BCE}}(\mathbf{S}, \mathbf{G}), \quad (23)$$

where \mathbf{G} denotes the ground-truth saliency map.

IV. EXPERIMENTS

A. Experimental Settings

Datasets. Two datasets are utilized for assessing the performance of SMN: HS-SOD [20] and our dataset, HSOD-BIT. HS-SOD comprises 60 HSIs with a spectral range of 380-780nm at intervals of 5nm, and a spatial resolution of 768×1024 pixels. For the purpose of evaluation, 48 HSIs are allocated for training, while 12 HSIs are reserved for testing. On the other hand, HSOD-BIT encompasses 319 HSIs, each possessing a spatial resolution of 1240×1680

pixels and a spectral range of 400-1000nm with intervals of 3nm. In HSOD-BIT, 255 HSIs are employed for training, while 64 HSIs are utilized for testing. Both datasets comprise RGB images as well as binarized ground-truth images that correspond to the respective HSIs.

Evaluation Metrics. The assessment of saliency map detection performance necessitates the utilization of established evaluation metrics. The metrics are delineated as follows:

Mean absolute error (MAE) quantifies the pixel-level discrepancy between the saliency map \mathbf{S} and the ground truth image \mathbf{G} :

$$\text{MAE} = \frac{1}{W \times H} \sum_{x=1}^W \sum_{y=1}^H |S_{xy} - G_{xy}|, \quad (24)$$

where W and H denote the width and height of the input image, respectively.

S-measure [38] (S_α) evaluates the structural fidelity of the saliency map \mathbf{S} and is defined as a weighted sum of region similarity S_r and object similarity S_o :

$$S_\alpha = \alpha \times S_r(\mathbf{S}, \mathbf{G}) + (1 - \alpha) \times S_o(\mathbf{S}, \mathbf{G}). \quad (25)$$

For the definitions of S_r and S_o , the reader is referred to [38]. We adopt $\alpha = 0.5$, as recommended in [38].

Precision-Recall (PR) curve [39] serves as a conventional metric for saliency evaluation. It is derived by thresholding the saliency map from 0 to 255 and subsequently computing precision and recall at each threshold level:

$$\text{Precision} = \frac{|\mathbf{B} \cap \mathbf{G}|}{|\mathbf{B}|}, \quad \text{Recall} = \frac{|\mathbf{B} \cap \mathbf{G}|}{|\mathbf{G}|}, \quad (26)$$

where \mathbf{B} and \mathbf{G} denote the binarized saliency maps and the ground truth, respectively.

F-measure [40] (F_β) is the harmonic mean of precision and recall, formulated as:

$$F_\beta = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}. \quad (27)$$

We employ the maximum F-measure, denoted as F_β^{\max} , for comparative analyses. The value of β^2 is set to 0.3, as suggested in [40].

Receiver Operating Characteristic (ROC) curve [39] is characterized by the true positive rate (TPR) and false positive rate (FPR):

$$\text{TPR} = \frac{|\mathbf{B} \cap \mathbf{G}|}{|\mathbf{G}|}, \quad \text{FPR} = \frac{|\mathbf{B} \cap \bar{\mathbf{G}}|}{|\bar{\mathbf{G}}|}. \quad (28)$$

Here, $\bar{\mathbf{G}}$ denotes the complement of the ground truth \mathbf{G} . Area Under Curve (AUC) is the total area under the ROC curve.

Correlation Coefficient (CC) [41] measures the statistical correlation between the saliency map \mathbf{S} and ground truth \mathbf{G} :

$$\text{CC} = \frac{\sigma(\mathbf{S}, \mathbf{G})}{\sigma(\mathbf{S}) \times \sigma(\mathbf{G})}, \quad (29)$$

where $\sigma(\mathbf{S}, \mathbf{G})$ is the covariance between \mathbf{S} and \mathbf{G} . Overall, a better HSOD saliency detector shall have a smaller MAE and larger other metrics.

Implementation Details. For the purpose of reducing memory cost and computational complexity, we performed downsampling on the original HSIs both spatially and spectrally. As a result, the HSIs were transformed into a spatial resolution of 224×224 pixels and consisted of 50 spectral channels. To augment the data, we employed horizontal flip and random crop techniques. In the low-frequency embedding phase, ResNet18 [42], Swin-tiny [43], and PVTv2-b1 [44] were utilized as base architectures, initialized with weights pre-trained on the ImageNet1k dataset. The models are denoted as SMN-R, SMN-S, and SMN-P, respectively. To implement cross-attention, we modified the neighborhood attention mechanism accordingly. The kernel size for the high-frequency and low-frequency attention heads was set to 13 and 9, respectively. Our model was trained on a single NVIDIA RTX 3090 GPU with an Intel XEON Gold 5218R CPU. Stochastic gradient descent (SGD) with a momentum optimizer was employed for training, spanning a total of 100 epochs. A warm-up and linear decay strategy was employed to calibrate the maximum learning rate to 2×10^{-2} (for Swin-tiny and PVTv2-b1, it was set to 7×10^{-3}). The batch size was configured to 5.

Competing methods. Itti's model [15] serves as the baseline model for HSOD. Initially, we compare our model with several conventional methods proposed by Liang *et al.* [16], namely spectral angular distance (SAD), spectral Euclidean distance (SED), and spectral grouping (SG). In order to compare with open-source state-of-the-art methods, we also include SUDF proposed by Imamoglu *et al.* [18] in the comparison. For the sake of fairness, SUDF retains the default parameter settings. Furthermore, we compare our SMN with two classical RGB-image-based SOD methods, BASNet [37] and U2Net [5], to validate the necessity of developing HSOD methods.

B. Results on HSOD-BIT

Quantitative Results. The quantitative comparison results on HSOD-BIT can be found in Table I. Regardless of the backbone network employed, our SMN consistently outperforms both traditional methods and SUDF. Specifically, our SMN-R achieves impressive scores of 0.039 for MAE, 0.869 for S_α , 0.854 for F_β^{\max} , 0.969 for AUC, and 0.849 for CC. These results surpass SUDF by 74.00%, 29.93%, 56.99%, 5.56%, and 26.53%, respectively. Utilizing other transformer-based backbones yields enhanced detection performance, validating the effectiveness of our SMN. Traditional methods heavily rely on manually designed low-level features and fail to effectively exploit the entire spectral information, thus limiting their ability to generate highly accurate saliency maps.

In comparison to U2Net, a classical RGB-image-based SOD method, our SMN-R exhibits superior performance in terms of F_β^{\max} , AUC, and CC, while slightly trailing behind U2Net in the MAE and S_α metrics. Employing Swin-tiny and PVTv2-b1 as backbones results in detection performance substantially superior to that of U2Net. Specifically, as shown in Table I, when employing Swin-tiny as the backbone (SMN-S), we observed a 4.7% increase in F_β^{\max} and a 3.1% increase in AUC compared to U2Net. Similarly, for the PVTv2-b1 backbone (SMN-P), the F_β^{\max} and AUC increased by 5.2% and 4.1% compared to U2Net on the HSOD-BIT dataset.

TABLE I
 QUANTITATIVE RESULTS ON HSOD-BIT AND HS-SOD DATASETS. ‘-R’: RESNET18 [42], ‘-S’: SWIN-TINY [43], ‘-P’: PVTv2-B1 [44].

| Datasets | HSOD-BIT | | | | | HS-SOD | | | | |
|--------------|--------------|--------------|--------------------|--------------|--------------|--------------|--------------|--------------------|--------------|--------------|
| | MAE ↓ | S_α ↑ | F_β^{\max} ↑ | AUC ↑ | CC ↑ | MAE ↓ | S_α ↑ | F_β^{\max} ↑ | AUC ↑ | CC ↑ |
| Itti [15] | 0.247 | 0.532 | 0.374 | 0.801 | 0.355 | 0.257 | 0.488 | 0.271 | 0.783 | 0.225 |
| SAD [16] | 0.203 | 0.552 | 0.390 | 0.830 | 0.397 | 0.203 | 0.500 | 0.244 | 0.778 | 0.223 |
| SED [16] | 0.130 | 0.500 | 0.343 | 0.753 | 0.303 | 0.132 | 0.470 | 0.291 | 0.793 | 0.201 |
| SG [16] | 0.182 | 0.543 | 0.338 | 0.791 | 0.370 | 0.196 | 0.530 | 0.274 | 0.808 | 0.268 |
| SUDF [18] | 0.150 | 0.685 | 0.544 | 0.918 | 0.671 | 0.242 | 0.498 | 0.275 | 0.723 | 0.250 |
| BASNet [37] | 0.040 | 0.849 | 0.779 | 0.919 | 0.785 | 0.071 | 0.743 | 0.605 | 0.843 | 0.625 |
| U2Net [5] | 0.034 | 0.870 | 0.829 | 0.942 | 0.830 | 0.076 | 0.734 | 0.617 | 0.854 | 0.631 |
| SMN-R (Ours) | 0.039 | 0.869 | 0.854 | 0.969 | 0.849 | 0.069 | 0.767 | 0.682 | 0.903 | 0.684 |
| SMN-S (Ours) | 0.032 | 0.891 | 0.868 | 0.971 | 0.870 | 0.079 | 0.737 | 0.659 | 0.899 | 0.635 |
| SMN-P (Ours) | 0.034 | 0.892 | 0.872 | 0.981 | 0.874 | 0.068 | 0.788 | 0.723 | 0.916 | 0.718 |

TABLE II
 QUANTITATIVE EFFICIENCY ANALYSIS. ‘-R’: RESNET18 [42], ‘-S’: SWIN-TINY [43], ‘-P’: PVTv2-B1 [44].

| Metrics | FLOPs (G) | #Params (M) | Speed (FPS) | F_β^{\max} ↑ |
|--------------|--------------|-------------|--------------|--------------------|
| SUDF [18] | 82.90 | 0.10 | 0.51 | 0.544 |
| BASNet [37] | 127.56 | 87.06 | 51.40 | 0.779 |
| U2Net [5] | 47.65 | 44.01 | 33.47 | 0.829 |
| SMN-R (Ours) | 14.58 | 7.27 | 35.91 | 0.854 |
| SMN-S (Ours) | 17.23 | 16.87 | 30.17 | 0.868 |
| SMN-P (Ours) | 14.76 | 10.23 | 32.68 | 0.872 |

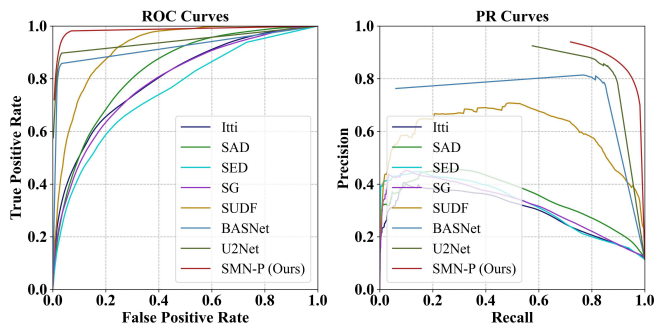


Fig. 3. Comparison of ROC and PR curves for multiple models on our HSOD-BIT dataset. Our SMN, represented by a red line, outperforms others.

These outcomes highlight the inefficacy of employing SOD methods after simply converting HSIs into false-color images. The presence of similar colors between the foreground and background in the false-color image poses a challenge in distinguishing between them. Conversely, our SMN fully utilizes spectral information, remaining unaffected by variations in illumination conditions, thereby enabling the detection of salient objects even in challenging scenarios.

Figure 3 provides a comparison of the ROC curves and PR curves between our SMN and other methods. Our SMN is denoted by the red line. Notably, the ROC curve of our SMN closely approaches the point (0, 1), while the PR curve is nearest to the point (1, 1) in comparison to the other methods. These observations indicate the superior performance of our SMN. The combined analysis of ROC curves, PR curves, and the accompanying numerical evaluation metrics serves to

validate the effectiveness of our SMN.

Efficiency Analysis. We conducted a comparative efficiency analysis of our SMN with other methods, including Floating Point Operations (FLOPs), number of parameters (#Params), and inference speed (FPS). The results are shown in Table II. It is worth noting that the spatial dimensions and the number of spectral channels for each method were kept at their respective default values. SUDF employs a CNN for feature extraction purposes only, followed by manifold learning and superpixel clustering. Consequently, it utilizes a relatively small number of parameters and exhibits a lower inference speed, at 0.1 M and 0.51 FPS, respectively. Moreover, due to the use of the entire HSI as input without spatial downsampling, SUDF incurs a high computational cost in terms of FLOPs. Our SMN demonstrates a reduction in FLOPs and an enhanced inference speed relative to SUDF. Moreover, in comparison to BASNet and U2Net, our approach significantly minimizes both the parameter and FLOPs, yet achieves commendable detection performance. Changing the backbone network results in a modest increase in FLOPs and the number of parameters, but does not significantly impact inference speed, while substantially enhancing detection performance. This demonstrates that SMN offers a good trade-off between computational efficiency, speed, and effectiveness.

Qualitative Results. The qualitative results obtained on HSOD-BIT are presented in Figure 4. In comparison to previous HSOD approaches, our proposed SMN demonstrates the ability to accurately and comprehensively detect salient objects. For instance, in scenes on rows 1, 3, and 5, some HSOD algorithms yield misleading saliency outcomes or struggle to detect salient objects effectively.

Furthermore, we compare our SMN with a well-known RGB-image-based SOD algorithm called U2Net [5]. Under normal circumstances, SMN achieves comparable detection performance to U2Net: both methods produce complete detection results with sharp edges. However, in scenes where the foreground and background colors are similar, SMN exhibits more precise edge delineation compared to U2Net. Moreover, in overexposed scenes, SMN showcases higher levels of detection accuracy and completeness relative to U2Net. This is attributed to the fact that U2Net relies solely on spatial or color information from the false-color image, rendering

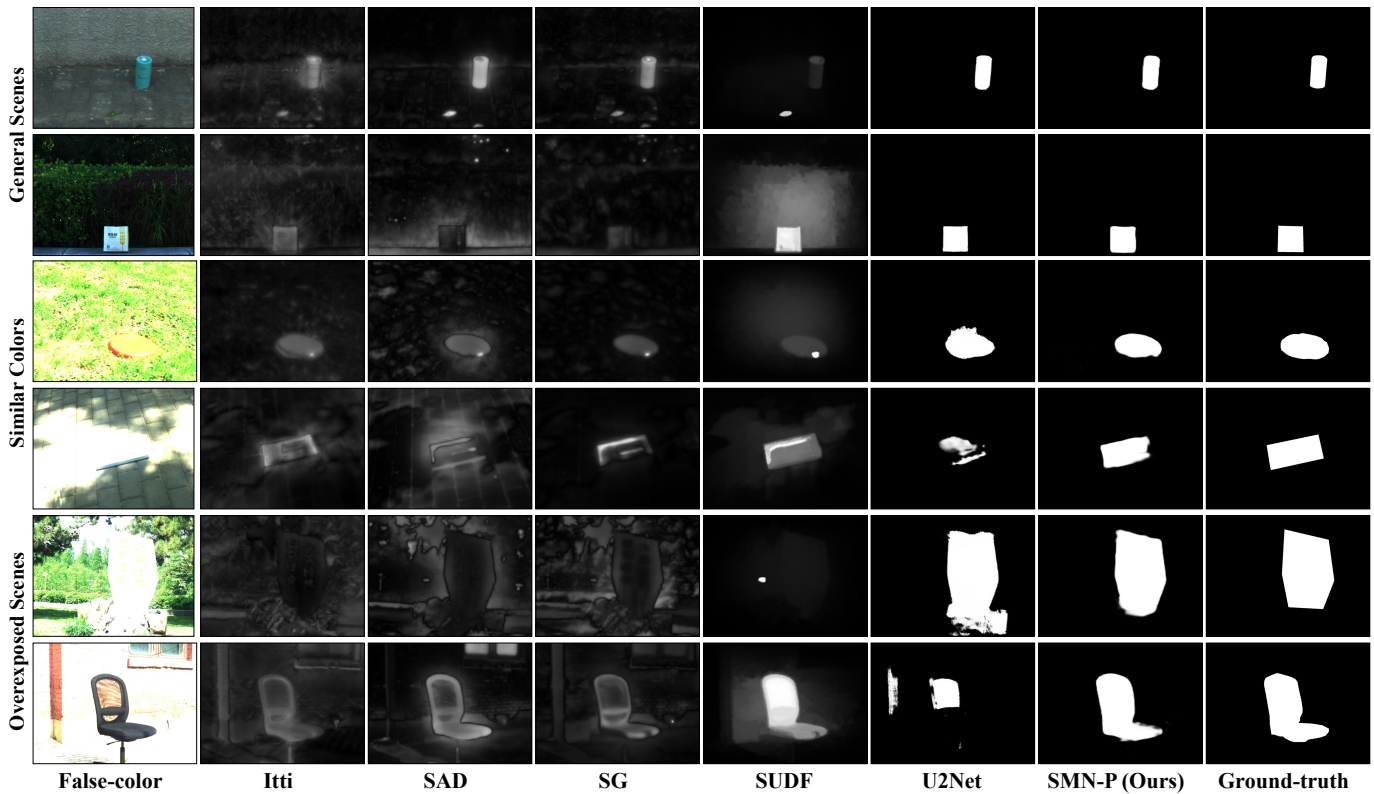


Fig. 4. Qualitative results on our HSOD-BIT dataset. SMN has better detection performance in similar colors and overexposed scenes.

it ineffective in distinguishing foreground from background in scenarios where color information is lacking. Moreover, U2Net does not utilize edge information to refine saliency, leading to erroneous detection results, as illustrated in the third and fifth rows of the scenes. In contrast, our SMN approaches the problem from a spectral perspective, extracting Spectral Saliency and Spectral Edge images separately, and combining them through a specially designed Mixed-frequency Attention mechanism to leverage their complementarity.

Visualization of Attention Features. The output features of the high-frequency attention head and the low-frequency attention head in the Mixed-frequency Attention are displayed in Figure 5. It is evident that the features generated by the high-frequency attention head are predominantly concentrated along the edges of salient objects. In contrast, the features produced by the low-frequency attention head are primarily focused on the salient objects themselves.

Visualization of SEO and SSG Output. The visualization of the Spectral Saliency and Spectral Edge maps can be observed in Figure 6. It is worth noting that the choice of kernel size significantly influences the resulting edge features. Employing a smaller gradient convolution kernel yields a more detailed edge image, as depicted in the leftmost edge map. Conversely, utilizing a larger gradient convolution kernel leads to a clearer overall contour of the object, as depicted in the rightmost edge map. Similarly, the saliency maps obtained from the upper layer pairs in the pyramid (with a smaller layer index c) typically encompass complete salient objects. In contrast, the saliency maps obtained from the lower layer pairs (with a

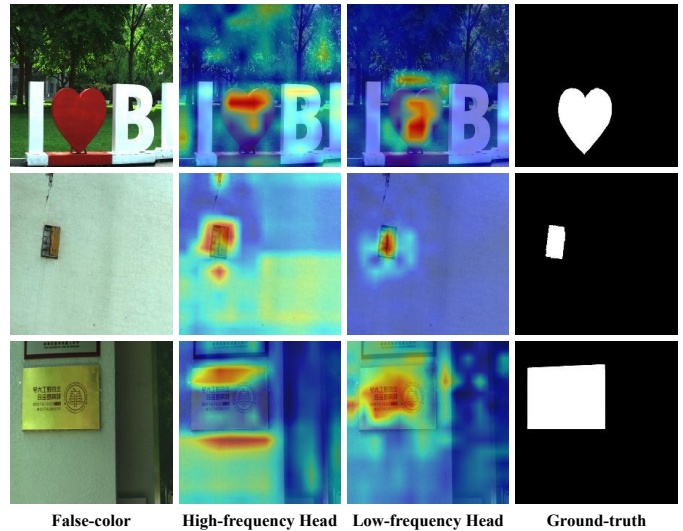


Fig. 5. Visualization of the output of features by the high-frequency attention head and the low-frequency attention head. The former attends to the edge of salient objects, while the latter focuses more on salient objects.

larger layer index c) offer greater accuracy in capturing salient regions.

C. Results on HS-SOD

Quantitative Results. The quantitative comparison results on the HS-SOD dataset are presented in Table I. Our SMN, irrespective of the backbone used, achieves notable performance

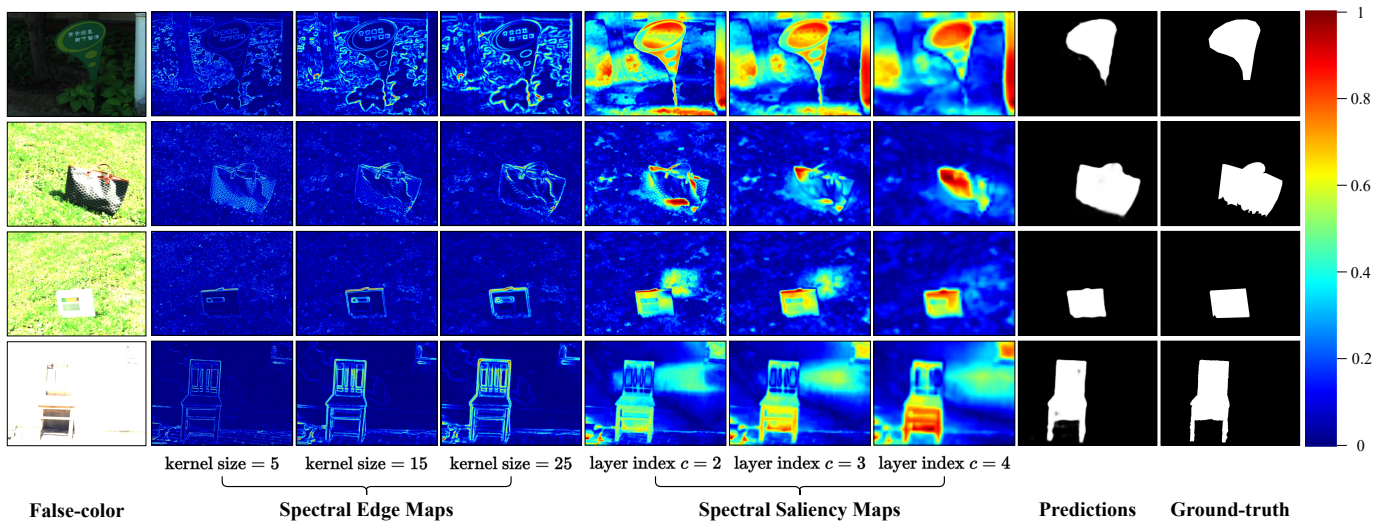


Fig. 6. Visualization of Spectral Edge and Spectral Saliency maps from SEO and SSG. The variation of gradient kernel sizes and the layer index c of the “center” pixel result in different Spectral Edge and Spectral Saliency maps, respectively.

scores, with SMN-P attaining MAE, S_α , F_β^{\max} , AUC, and CC values of 0.068, 0.788, 0.723, 0.916, and 0.718, respectively. Notably, our method outperforms traditional methods and SUDF across all evaluation metrics on this dataset. Compared to RGB-image-based methods, SMN-S slightly lags behind in two metrics, namely MAE and S_α . However, it outperforms both BASNet and U2Net in the remaining three evaluation metrics. SMN-S outperforms U2Net by increasing F_β^{\max} by 6.8% and AUC by 5.3%. Similarly, SMN-P outperforms U2Net by increasing F_β^{\max} by 17.2% and AUC by 7.2%. These results underscore the overall efficacy and competitiveness of our proposed SMN on the HS-SOD dataset.

The comparison of ROC and PR curves between our SMN and other methods on the HS-SOD dataset can be observed in Figure 7. The ROC curve of SMN, depicted by the red line, demonstrates its proximity to the point (0, 1), indicating a clear advantage over the other methods. Moreover, its advantage in the PR curve is also obvious. By considering the ROC curves, PR curves, and various evaluation metrics, our SMN showcases effectiveness in the context of the HS-SOD dataset.

Qualitative Results. The qualitative results on the HS-SOD dataset are presented in Figure 8. It can be observed that conventional methods and SUDF exhibit numerous errors and incompleteness in their saliency detection outputs. In a typical scenario, when compared to U2Net, our SMN demonstrates higher accuracy in detecting objects such as tree trunks and street lamps. This is attributed to the fact that U2Net solely relies on color information for salient object detection and struggles to differentiate objects with similar colors accurately. Conversely, SMN leverages spectral information derived from material properties, enabling it to better distinguish objects with similar colors. However, in more complex scenes, both SMN and U2Net exhibit errors in their detection results, suggesting the challenges associated with accurate detection. Regarding small objects, our SMN model exhibits a more comprehensive detection result compared to U2Net. For in-

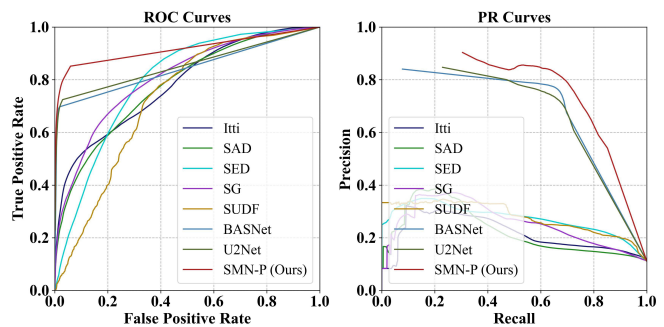


Fig. 7. Comparison of ROC and PR curves for multiple models on HS-SOD dataset. SMN, represented by a red line, demonstrates a clear advantage in both ROC and PR curves.

stance, in the small object scene, SMN successfully identifies the seated person’s back as a salient object, whereas U2Net only detects a portion of the person’s head. On the other hand, in the large object scene, both SMN and U2Net face challenges in fully detecting the target objects. In such cases, the performance of both methods is limited.

D. Ablation Study

We conducted ablation studies on our HSOD-BIT dataset, choosing PVTv2-b1 [44] as the backbone network for the low-frequency embedding.

Hyperparameter Analysis. As previously mentioned, the kernel size of the neighborhood attention mechanism employed in the high-frequency and low-frequency heads differs due to the distinct input and objective of these heads. Hence, we conducted a hyperparameter analysis on the kernel sizes, as well as the number of attention heads. The results, depicted in Figure 9, highlight the significant impact of these hyperparameters on the model’s detection performance. For instance, let us consider the evaluation metrics AUC and F_β^{\max} . As the kernel size of high-frequency attention increases, the model’s

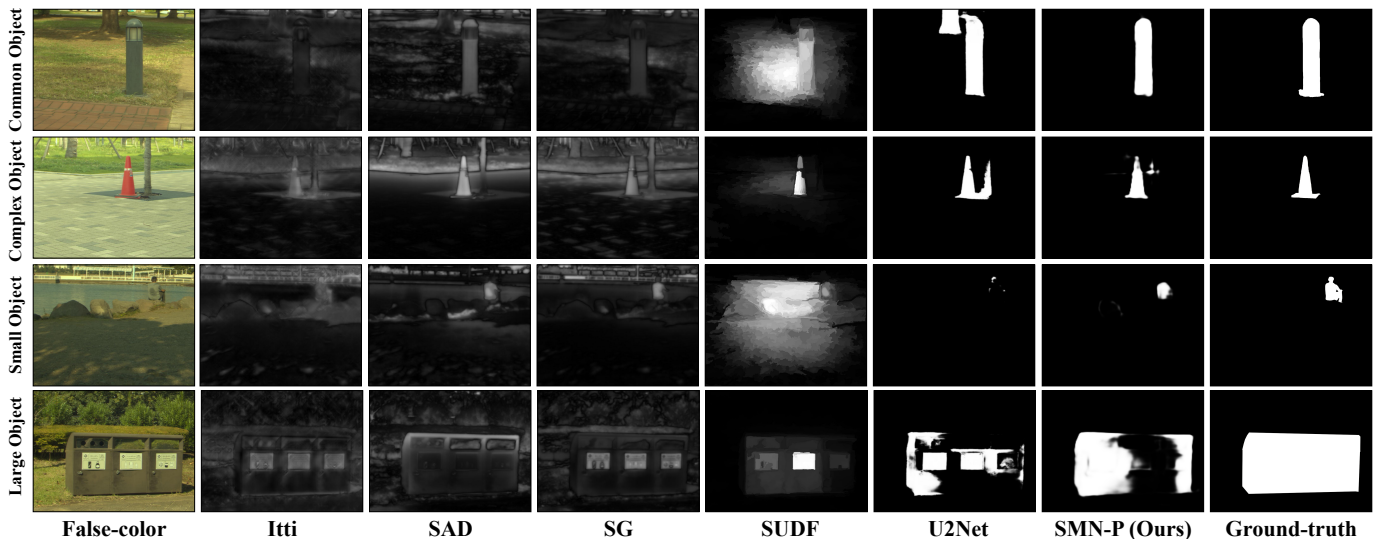


Fig. 8. Qualitative Results on HS-SOD dataset. SMN outperforms other methods and is most similar to the ground-truth.

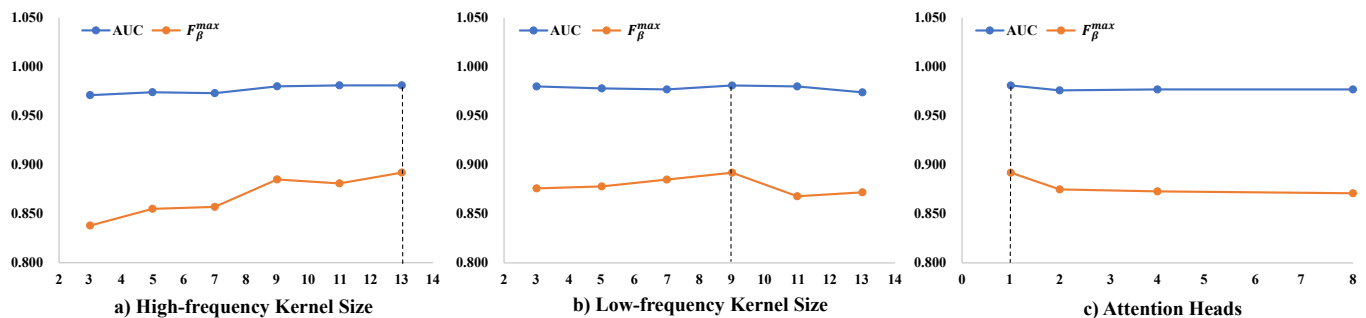


Fig. 9. Hyperparameter analysis of the kernel sizes and number of heads in the neighborhood attention mechanism.

detection performance gradually improves, reaching its peak when the kernel size is 13. Further increasing the kernel size may yield better results; however, it is important to note that the neighborhood attention mechanism currently supports a maximum kernel size of 13, limiting any further increase.

Regarding the low-frequency branch, the detection performance of the model achieves its highest value when the kernel size is set to 9, taking into account both evaluation metrics. Furthermore, increasing the number of attention heads has a noticeable negative impact on the model’s detection performance. Based on these findings, we have determined the optimal values for the three hyperparameters: the kernel size for the high-frequency attention is set to 13, the kernel size for the low-frequency attention is set to 9, and each attention mechanism employs a single attention head.

Comparison with Inputting RGB Images. In order to assess the importance of using HSI as input, we conducted an experiment where we removed two modules specifically designed for HSI, namely SSG and SEO, and directly input false-color images into the SMN. It is worth noting that the hyperparameters of the network remained unchanged throughout this experiment. As illustrated in Table III, the performance of the modified SMN model, measured in terms of F_{β}^{\max} and AUC, yielded values of 0.887 and 0.978, respectively. These values

TABLE III
ABLATION STUDY OF INPUT DATA.

| False-color | HSI | Spec. Edge | Spec. Sal. | $F_{\beta}^{\max} \uparrow$ | AUC \uparrow |
|-------------|-----|------------|------------|-----------------------------|----------------|
| ✓ | ✗ | ✗ | ✗ | 0.887 | 0.978 |
| ✗ | ✓ | ✗ | ✗ | 0.867 | 0.967 |
| ✗ | ✓ | ✓ | ✗ | 0.877 | 0.972 |
| ✗ | ✓ | ✗ | ✓ | 0.881 | 0.979 |
| ✗ | ✓ | ✓ | ✓ | 0.892 | 0.981 |

were found to be lower compared to the complete SMN model, which achieved scores of 0.892 and 0.981 in the same metrics. This outcome clearly indicates that the simple conversion of HSI to false-color images is less effective in the context of salient object detection, emphasizing the necessity of utilizing HSI as input for achieving superior performance.

Usefulness of Two Plug-and-play Operators. To assess the effectiveness of the plug-and-play modules, SEO and SSG, we conducted experiments where we removed each module individually and compared the results with the baseline.

The baseline experiment involved inputting the complete HSI into the SMN without any modifications. When both SEO and SSG modules were removed, the first convolutional layer in the frequency-specified embeddings was randomly

TABLE IV
THE EFFECTIVENESS OF MIXED-FREQUENCY ATTENTION MODULE,
SHALLOW FEATURE, AND HIGH-FREQUENCY INFORMATION.

| Models | MAE ↓ | F_{β}^{\max} ↑ | S_{α} ↑ | AUC ↑ | CC ↑ |
|--------------------------|--------------|----------------------|----------------|--------------|--------------|
| SMN <i>w/o</i> MA | 0.036 | 0.862 | 0.878 | 0.978 | 0.863 |
| SMN <i>w/o</i> Sha. Fea. | 0.040 | 0.848 | 0.868 | 0.970 | 0.845 |
| SMN <i>w/o</i> HF | 0.033 | 0.871 | 0.887 | 0.976 | 0.871 |
| SMN | 0.034 | 0.892 | 0.872 | 0.981 | 0.874 |

initialized, and the input channels were changed to 50. This configuration resulted in the poorest detection performance, with an F_{β}^{\max} score of only 0.867 and an AUC of 0.967. This outcome highlights the usefulness of the SEO and SSG modules in improving the detection performance of the SMN, underscoring their significance in the context of HSOD.

Effect of SEO. By incorporating the SEO module, we made modifications to the inputs of the high-frequency and low-frequency embeddings. Specifically, the Spectral Edge map was used as the input for the high-frequency embedding, while the original HSI was retained as the input for the low-frequency embedding. A comparison between the second row and the third row of Table III reveals notable improvements in F_{β}^{\max} and AUC, indicating a significant enhancement in the detection performance. These results serve as evidence supporting the effectiveness of the SEO module.

Efficacy of SSG. Upon integrating the SSG module into the baseline configuration, the input for SMN consists of the complete HSI and the Spectral Saliency map. The results depicted in Table III exhibit noticeable improvements in both the F_{β}^{\max} and AUC metrics, thereby affirming the effectiveness of the SSG module. Furthermore, as the Spectral Saliency map provides valuable insights into the approximate location of salient objects, it serves as a crucial information source for SMN’s saliency detection. Consequently, the inclusion of the SSG module yields more substantial enhancements in the detection performance compared to solely employing SEO.

Impact of SEO and SSG. The final row of Table III demonstrates that the combined utilization of the SEO and SSG modules, which convert the HSI into edge images and saliency maps, respectively, produces the most remarkable detection performance. Notably, the simultaneous application of both modules yields a more substantial enhancement in detection performance compared to employing either module individually. This outcome can be attributed to the fact that the SEO and SSG modules effectively transform the HSI into edge images and saliency maps, respectively. These transformed representations provide more accurate and suitable high-frequency and low-frequency information for SMN, aligning with the requirements of our specially designed model.

Effect of Mixed-frequency Attention Module. We investigate the impact of the Mixed-frequency Attention (MA) module. An alternative feature fusion approach involved concatenating deep edge information and saliency information along the channel dimension while also modifying the input channel number of the first convolutional layer in the Saliency-edge-aware Decoder. This experimental setup is denoted as SMN *w/o* MA.

By comparing the results in the first and last rows of Table IV, it becomes evident that the inclusion of MA has a significant positive effect on the detection performance of the model. MA facilitates the self-refinement of low-frequency information through the utilization of a self-attention mechanism, enabling it to concentrate more on salient objects. Furthermore, the cross-attention mechanism promotes interaction between high and low-frequency features, leading to the generation of more accurate low-frequency features within the constraints imposed by high-frequency information.

Effect of Shallow Feature. To investigate the influence of shallow features in the Saliency-edge-aware Decoder, we conducted an experiment where these features were eliminated, resulting in a conventional decoder. This experimental setup is referred to as SMN *w/o* Sha. Fea. Upon examining the results presented in Table IV, it becomes apparent that the inclusion of shallow features during the decoding process significantly enhances the detection performance. As the model progresses deeper into the network, the spatial dimensions of the features gradually decrease, leading to a loss of fine details. When decoding is performed solely based on these deep features, the resulting outcomes become less precise. Therefore, integrating shallow features from the encoder at the decoding stage is crucial to compensate for the loss of intricate information and improve the overall detection performance.

Impact of High-frequency Information. To investigate the role of high-frequency information in salient object detection, we conducted an experiment where we removed the high-frequency inputs and solely relied on low-frequency information. This experimental setup involved generating a Spectral Saliency map from the HSI using the SSG module and converting it into low-frequency features through low-frequency embeddings. Subsequently, these features underwent self-attention in the MA for self-refinement and were decoded to obtain the final saliency map without including shallow edge information as input to the decoder. This experiment is denoted as SMN *w/o* HF. By comparing the last two lines in Table IV, it becomes evident that including high-frequency information is crucial for achieving robust salient object detection performance in the SMN model. The results demonstrate the necessity of incorporating inputs from both high and low frequencies for effective detection.

E. Extensive Experiment on RGBT SOD

Thermal images possess the capability to effectively capture temperature information pertaining to objects within a given scene. In these images, objects exhibiting higher temperatures exhibit greater intensity, thus distinguishing themselves. This particular characteristic bears a resemblance to our Spectral Saliency map. As a result, our proposed SMN is employed in RGBT datasets with the purpose of assessing and confirming the generalizability of our proposed methodology.

Experimental Settings. Given the alterations in the task and dataset, we adjusted the batch size to 32 and set the maximum learning rate to 1×10^{-2} . The backbone network is PVTv2-b1. **Datasets.** Our training and test sets are consistent with the dataset used by Tu *et al.* [45]. To provide a comprehensive

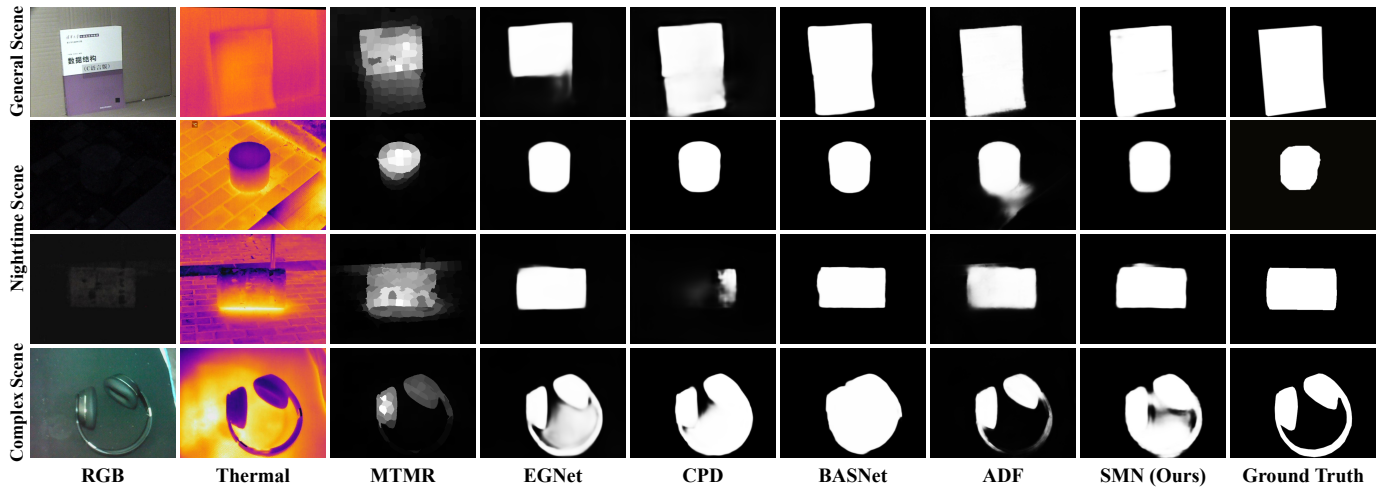


Fig. 10. Qualitative results of RGB T SOD. Our SMN achieves satisfactory detection results.

TABLE V
RESULTS FOR RGB T SOD ON VT821, VT1000 AND VT5000 DATASETS.

| Methods | VT821 | | VT1000 | | VT5000 | |
|-------------|-----------------------------|------------------|-----------------------------|------------------|-----------------------------|------------------|
| | $F_{\beta}^{\max} \uparrow$ | MAE \downarrow | $F_{\beta}^{\max} \uparrow$ | MAE \downarrow | $F_{\beta}^{\max} \uparrow$ | MAE \downarrow |
| MTMR [46] | 0.747 | 0.108 | 0.754 | 0.119 | 0.662 | 0.115 |
| EGNet [47] | 0.795 | 0.063 | 0.917 | 0.033 | 0.839 | 0.051 |
| CPD [48] | 0.786 | 0.079 | 0.914 | 0.031 | 0.847 | 0.047 |
| BASNet [37] | 0.803 | 0.067 | 0.913 | 0.030 | 0.820 | 0.055 |
| ADF [49] | 0.804 | 0.077 | 0.923 | 0.034 | 0.863 | 0.048 |
| SMN (Ours) | 0.831 | 0.043 | 0.911 | 0.027 | 0.908 | 0.044 |

comparison, we compare the performance of SMN with several existing methods, including MTMR [46], EGNet [47], CPD [48], BASNet [37], and ADF [49]. Evaluation of the methods is conducted using two metrics: F_{β}^{\max} and MAE.

Quantitative Results. As depicted in Table V, the SMN demonstrates commendable detection outcomes, exhibiting the lowest MAE values across all three datasets, namely 0.043, 0.027, and 0.044, respectively. Moreover, on the F_{β}^{\max} metric, SMN outperforms ADF on both the VT821 and VT5000 datasets and only lags behind ADF by 0.012 on the VT1000 dataset. During the RGB T SOD experiment, the SSG and SEO modules are excluded. These modules, integral to processing hyperspectral data, extract edge and saliency information from a spectral standpoint. Nonetheless, in the RGB T SOD experiment, solely RGB images are employed for edge extraction, leading to an inherent loss of information. Additionally, while our Spectral Saliency maps encompass a collection of spectral saliency images, there exists only a single thermal image, thereby providing comparatively less information. Despite these, SMN’s detection performance remains robust, amply demonstrating the effectiveness and superiority of our proposed SMN.

Qualitative Results. The qualitative results of RGB T SOD are visually presented in Figure 10. Our proposed methods have demonstrated promising outcomes in terms of salient object detection. When compared to MTMR, SMN exhibits enhanced

accuracy and completeness in detecting salient objects across diverse scenes. In the general scene, SMN achieves detection results better than MTMR and EGNet. In nighttime scenes, SMN outperforms MTMR, CPD, and ADF by producing more precise detection outcomes with sharper edges. Nonetheless, in the complex scene, SMN exhibits limitations in effectively detecting finer details of head-worn headphones, resulting in relatively weaker performance when compared to ADF.

V. CONCLUSION

In this study, we introduce a novel lightweight model, Spectrum-driven Mixed-frequency Network (SMN), for hyperspectral salient object detection. Our approach is motivated by the insight that spectral information can be leveraged to extract features with two distinct frequencies. To this end, we develop two plug-and-play operators, namely the Spectral Saliency Generator and the Spectral Edge Operator. Furthermore, we design a customized Mixed-frequency Attention module that effectively utilizes the complementarity of these features to generate saliency maps with high-fidelity edges. Experiment results demonstrate our SMN’s superiority to state-of-the-art HSOD methods.

Although our method currently surpasses those based on RGB images, the advantage is not substantial. In the future, we plan to construct datasets that more effectively highlight the benefits of utilizing hyperspectral information for saliency detection. Additionally, we aim to further reduce the model size and enhance its speed, facilitating deployment on computation and memory-limited devices.

VI. ACKNOWLEDGEMENT

This work was financially supported by the National Key Scientific Instrument and Equipment Development Project of China (No. 61527802), the National Natural Science Foundation of China (No. 62101032), the Postdoctoral Science Foundation of China (Nos. 2021M690015, 2022T150050), and Beijing Institute of Technology Research Fund Program for Young Scholars (No. 3040011182111).

REFERENCES

- [1] C. Huang, T. Xu, Y. Zhang, C. Pan, J. Hao, and X. Li, "Salient object detection on hyperspectral images in wireless network using cnn and saliency optimization," *Ad Hoc Networks*, vol. 112, p. 102369, 2021.
- [2] Y. Cao, J. Zhang, Q. Tian, L. Zhuo, and Q. Zhou, "Salient target detection in hyperspectral images using spectral saliency," in *ChinaSIP*, 2015, pp. 1086–1090.
- [3] B. Arad and O. Ben-Shahar, "Sparse recovery of hyperspectral signal from natural rgb images," in *ECCV*, 2016, pp. 19–34.
- [4] Z. Tianyu and J. Xu, "Hyperspectral remote sensing image segmentation based on the fuzzy deep convolutional neural network," in *CISP-BMEI*, 2020, pp. 181–186.
- [5] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern Recognition*, vol. 106, p. 107404, 2020.
- [6] A. Borji, "What is a salient object? a dataset and a baseline model for salient object detection," *IEEE Transactions on Image Processing*, vol. 24, no. 2, pp. 742–756, 2015.
- [7] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *CVPR*, June 2015.
- [8] J. Zhang, Y. Dai, and F. Porikli, "Deep salient object detection by integrating multi-level cues," in *WACV*, 2017, pp. 1–10.
- [9] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *ICCV*, October 2021, pp. 4722–4732.
- [10] J. Zhang, J. Xie, N. Barnes, and P. Li, "Learning generative vision transformer with energy-based latent space for saliency prediction," in *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [11] M. Zhuge, D.-P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao, "Salient object detection via integrity learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3738–3752, 2023.
- [12] J. Liang, J. Zhou, L. Tong, X. Bai, and B. Wang, "Material based salient object detection from hyperspectral images," *Pattern Recognition*, vol. 76, pp. 476–490, 2018.
- [13] S. Sun, J. Liu, X. Chen, W. Li, and H. Li, "Hyperspectral anomaly detection with tensor average rank and piecewise smoothness constraints," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2022.
- [14] C. Jänicke, A. Okujeni, S. Cooper, M. Clark, P. Hostert, and S. van der Linden, "Brightness gradient-corrected hyperspectral image mosaics for fractional vegetation cover mapping in northern california," *Remote Sensing Letters*, vol. 11, no. 1, pp. 1–10, 2020.
- [15] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.
- [16] J. Liang, J. Zhou, X. Bai, and Y. Qian, "Salient object detection in hyperspectral imagery," in *2013 IEEE International Conference on Image Processing*, 2013, pp. 2393–2397.
- [17] S. Le Moan, A. Mansouri, J. Y. Hardeberg, and Y. Voisin, "Saliency for spectral image analysis," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 6, pp. 2472–2479, 2013.
- [18] N. İmamoğlu, G. Ding, Y. Fang, A. Kanazaki, T. Kouyama, and R. Nakamura, "Salient object detection on hyperspectral images using features learned from unsupervised segmentation task," in *ICASSP*, 2019, pp. 2192–2196.
- [19] A. Hassani, S. Walton, J. Li, S. Li, and H. Shi, "Neighborhood attention transformer," in *CVPR*, June 2023, pp. 6185–6194.
- [20] N. Imamoglu, Y. Oishi, X. Zhang, G. Ding, Y. Fang, T. Kouyama, and R. Nakamura, "Hyperspectral image dataset for benchmarking on salient object detection," in *QoMEX*, 2018, pp. 1–3.
- [21] P. L. Rosin, "A simple method for detecting salient regions," *Pattern recognition*, vol. 42, no. 11, pp. 2363–2371, 2009.
- [22] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 73–80.
- [23] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, "Weakly-supervised salient object detection via scribble annotations," in *CVPR*, June 2020.
- [24] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Computational visual media*, vol. 5, pp. 117–150, 2019.
- [25] Z. Yao and L. Wang, "Boundary information progressive guidance network for salient object detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 4236–4249, 2022.
- [26] T. Wilson, S. Rogers, and M. Kabrisky, "Perceptual-based image fusion for hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 35, no. 4, pp. 1007–1017, 1997.
- [27] H. Zhang, H. Peng, M. D. Fairchild, and E. D. Montag, "Hyperspectral image visualization based on a human visual model," in *Human Vision and Electronic Imaging XIII*, vol. 6806, 2008, p. 68060N.
- [28] V. Mnih, N. Heess, A. Graves, and k. kavukcuoglu, "Recurrent models of visual attention," in *Advances in Neural Information Processing Systems*, vol. 27, 2014, pp. 2204–2212.
- [29] W. Yin, H. Schütze, B. Xiang, and B. Zhou, "ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 259–272, 06 2016.
- [30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 6000–6010.
- [31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.
- [32] H. Zhu, X. Sun, Y. Li, K. Ma, S. K. Zhou, and Y. Zheng, "Dftr: Depth-supervised fusion transformer for salient object detection," *arXiv preprint arXiv:2203.06429*, 2022.
- [33] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4486–4497, 2022.
- [34] Y. Yang, Q. Qin, Y. Luo, Y. Liu, Q. Zhang, and J. Han, "Bi-directional progressive guidance network for rgb-d salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5346–5360, 2022.
- [35] X. Li, Y. Xu, L. Ma, Z. Yang, Z. Huang, H. Hong, and J. Tian, "Multi-source weakly supervised salient object detection via boosting weak-annotation source and constraining object structure," *Digital Signal Processing*, vol. 126, p. 103461, 2022.
- [36] Z. Su, W. Liu, Z. Yu, D. Hu, Q. Liao, Q. Tian, M. Pietikainen, and L. Liu, "Pixel difference networks for efficient edge detection," in *ICCV*, 2021, pp. 5117–5127.
- [37] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *CVPR*, 2019.
- [38] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4548–4557.
- [39] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.
- [40] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgb-d salient object detection: A benchmark and algorithms," in *ECCV*. Springer, 2014, pp. 92–109.
- [41] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2019.
- [42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, June 2016.
- [43] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, October 2021, pp. 10012–10022.
- [44] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvtv2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 1–10, 2022.
- [45] R. Cong, K. Zhang, C. Zhang, F. Zheng, Y. Zhao, Q. Huang, and S. Kwong, "Does thermal really always matter for rgb-t salient object detection?" *IEEE Transactions on Multimedia*, pp. 1–12, 2022.
- [46] G. Wang, C. Li, Y. Ma, A. Zheng, J. Tang, and B. Luo, "Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach," in *Image and Graphics Technologies and Applications, IGTA 2018*. Springer, 2018, pp. 359–369.
- [47] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in *ICCV*, 2019.
- [48] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *CVPR*, 2019.
- [49] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "Rgb-t salient object detection: A large-scale dataset and benchmark," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.