Cross-Model Nested Fusion Network for Salient Object Detection in Optical Remote Sensing Images

Mingzhu Xu, *Member, IEEE*, Sen Wang, Yupeng Hu, *Member, IEEE*, Haoyu Tang, *Member, IEEE*, Runmin Cong, *Senior Member, IEEE*, and Liqiang Nie, *Senior Member, IEEE*

***Abstract*—Recently, salient object detection (SOD) in optical remote sensing images, dubbed ORSI-SOD, has attracted increasing research interest. Although deep-based models have achieved impressive performance, several limitations remain: a single image contains multiple objects with varying scales, complex topological structures, and background interference. These unresolved issues render ORSI-SOD a challenging task. To address these challenges, we introduce a distinctive cross-model nested fusion network (CMNFNet), which leverages heterogeneous features to increase the performance of ORSI-SOD. Specifically, the proposed model comprises two heterogeneous encoders, a conventional CNN-based encoder that can model local features, and a specially designed graph convolutional network (GCN)-based encoder with local and global receptive fields that can model local and global features simultaneously. To effectively differentiate between multiple salient objects of different sizes or complex topological structures within an image, we project the image into two different graphs with different receptive fields and conduct message passing through two parallel graph convolutions. Finally, the heterogeneous features extracted from the two encoders are fused in the well-designed attention enhanced cross model nested fusion module (AECMNFM). This module is meticulously crafted to integrate features progressively, allowing the model to adaptively eliminate background interference while simultaneously refining the feature representations. We conducted comprehensive experimental analyzes on benchmark datasets. The results demonstrate the superiority of our CMNFNet over 16 state-of-the-art (SOTA) models.**

***Index Terms*—Cross-model nested fusion, graph convolution network, optical remote sensing images (ORSIs), salient object detection (SOD).**

I. Introduction

VISUAL salient object detection (SOD) seeks to precisely distinguish and uniformly segment a visually prominent object from its surrounding background [1], [2], [3]. For the natural scene images (NSIs), remarkable advancements have been made in the domain of SOD, owing to advancements in deep convolution neural networks [4], [5], [6], [7]. Recently, researchers have become increasingly interested in extending SOD techniques from NSIs to optical remote sensing images (ORSIs), a technique referred to as ORSI-SOD [8], [9], [10]. ORSI-SOD is designed to precisely pinpoint visually appealing objects in ORSIs, independent of their categories. It can provide prior information for diverse subsequent tasks in ORSIs analysis and enhance the performance of the corresponding tasks, including super-resolution [11], scene classification [12], target extraction [13], and semantic segmentation [14]. ORSI-SOD faces unique obstacles that are not encountered in the realm of NSIs, thereby hindering effective detection. Attempts to directly apply existing NSI-SOD techniques to ORSI-SOD frequently yield unsatisfactory outcomes, as evident in column (f) of Fig. 1. This disparity is due to the distinct imaging methods, which involve capturing images from high altitudes using optical remote sensing satellites or aircraft [15]. ORSIs typically encompass vast areas, comprising various objects and intricate backgrounds. As a result, ORSI-SOD poses a more formidable challenge, necessitating a deeper investigation.

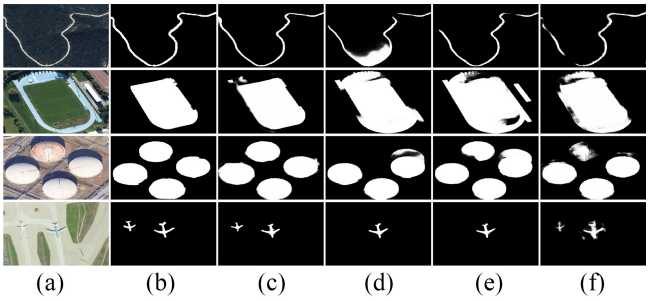

Fig. 1. Visual comparative analysis between our method and several advanced SOD models under classic and challenging ORSIs. (a) Original ORSIs. (b) Groundtruth. (c) Ours method. (d) ACCo-V (ORSI-SOD Model) [16]. (e) Corr-V (ORSI-SOD Model) [17]. (f) Gated (NSI-SOD Model) [18].

Recently, by tailoring to the unique properties of ORSIs, researchers have proposed numerous novel insights and crafted exceptional algorithms, which have significantly promoted the development of ORSI-SOD and delivered remarkable performance improvements [8], [9], [10], [16], [19], [20], [21], [22], [23]. Some works have conducted in-depth analyses on salient objects with variable scales in ORSIs and explored several multiscale feature extraction and integration approaches,

Received 22 March 2025; revised 16 May 2025; accepted 17 May 2025. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62206157, and in part by the Natural Science Foundation of Shandong Province under Grant ZR2022QF047. This article was recommended by Associate Editor A. Ferreira de Loza. *(Corresponding author: Yupeng Hu.)*

Mingzhu Xu, Sen Wang, Yupeng Hu, and Haoyu Tang are with the School of Software, Shandong University, Jinan 250101, Shandong, China (e-mail: xumingzhu@sdu.edu.cn; samwangyl1010@gmail.com; huyupeng@sdu.edu.cn; tanghao258@sdu.edu.cn).

Runmin Cong is with the School of Control Science and Engineering, Shandong University, Jinan 250061, Shandong, China (e-mail: rmcong@sdu.edu.cn).

Liqiang Nie is with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, Guangdong, China (e-mail: nieliqiang@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2025.3571913.

Digital Object Identifier 10.1109/TCYB.2025.3571913

aiming to perceive salient objects with various scales [8], [9], [10], [20], [21]. Inspired by the edge-aided SOD in NSIs, many works deeply mine the edge/boundary-guided CNN architecture to refine the coarse boundary of salient objects with complicated edges [10], [19], [22], [23]. The adjacent cross-layer features are also explored to adaptively complement the local details and global information, enhancing the integrity of the salient regions [16]. The hybrid features borrowed from CNNs and Transformers are grafted in a one-stream network, aiming to alleviate background interference [23]. Foreground/background decoupling has also been exploited to enhance context [24] and repair local attention loss [25]. These advanced ORSI-SOD models have attained notable performance.

However, the current methods still encounter several challenges when processing ORSIs. First, objects within ORSIs exhibit significant variations in size within the same image and the number of salient objects varies greatly across different images (e.g., rows 3 and 4 of Fig. 1). Second, the objects in ORSIs possess complex topological structures. There are numerous objects belonging to diverse categories, such as islands, rivers, airplanes, cars, houses, ships, etc., which have different topological structures and greatly increase the difficulty of detection (e.g., row 1 of Fig. 1). Moreover, ORSIs often feature complex and intricate backgrounds. This arises from the imaging technique employed in ORSIs, wherein all objects presented within the captured area appear equally. Consequently, ORSIs exhibit rich but complex backgrounds(e.g., rows 2 and 3 of Fig. 1).

To address these problems, we innovatively devised a cross-model nested fusion network (CMNFNet), which exploits heterogeneous features to increase performance in the ORSI-SOD task. Specifically, our proposed CMNFNet comprises two heterogeneous encoder types: 1) a conventional CNN-based encoder that can model local pattern features, and 2) a graph convolutional network (GCN)-based encoder with local and global receptive fields that can model local and global pattern features simultaneously. To perceive the salient objects with varying scales or complex topological structures within a single image, we project the image into two different graphs with different receptive fields and conduct message passing through two parallel graph convolutions. It allows our network to detect and identify salient objects precisely regardless of their varying scales or complex topological structures. Finally, the heterogeneous features extracted from the five stages of the two encoders are fused in the well-designed attention enhancement cross model nested fusion module (AECMNFM). Directly fusing the features from different encoders may lead to low performance. Our nested fusion model can complete the features in a progressive way, and can better combine global and local features adaptively.

The following offers a summation of our contributions.

1) We innovatively devise a CMNFNet for ORSI-SOD. Unlike existing models that rely on a single encoder or directly fuse off-the-shelf CNN and Transformer encoders, CMNFNet employs two heterogeneous encoders: a CNN and a custom-designed GCN-based encoder. It progressively fuses their features through a novel nested fusion strategy, effectively complementing heterogeneous representations without relying on existing direct fusion methods.
2) We propose a novel graph-based convolution subnetwork (Encoder-GCN) as an auxiliary encoder. Unlike traditional CNN or Transformer encoders, which can model only local or global context within each layer, our method employs dual parallel graph convolutions in distinct semantic spaces with varying receptive fields at each GCN layer, enabling joint modeling of both the local and global context. This facilitates the multiscale perception of complex salient objects (CSOs).
3) We propose a novel AECMNFM. Unlike traditional dual-/single-stream fusion methods, which tend to cause mutual interference between heterogeneous features, our approach progressively integrates these features in a complementary manner. This process is further enhanced by a hybrid attention mechanism that selectively emphasizes salient information while effectively suppressing background noise.
4) We conduct thorough experiments across three challenging datasets to evaluate the overall performance and the component effectiveness. Our method outperforms 16 advanced models, firmly demonstrating its superiority in ORSI-SOD.

II. Related Works

In this section, we provide a concise overview of the SOD models in both NSIs and ORSIs.

A. Salient Object Detection in NSIs

SOD was initially introduced to identify objects that attract attention within NSIs. In the preliminary models, the hand-crafted features are first well-designed to represent the discriminative spatial features of the salient objects and then learn to seamlessly integrate these saliency cues in an unsupervised manner, ultimately yielding saliency outcomes. The conventional methods encompass various approaches, including saliency tree [26], Bayesian inference [27], random walk ranking [28], [29], [30], structured matrix decomposition [31], sparse reconstruction [32], probability graphs [33], [34], etc. All these traditional methods have introduced a plethora of innovative viewpoints, offering new ideas for the further exploration of deep CNN-based methods.

In recent years, deep CNNs have been widely studied for SOD, resulting in numerous advanced deep CNN models to increase NSI-SOD performance. In the original deep CNN models, pyramidal features and dilated convolution are commonly utilized to extract multiscale context [35]. To strengthen the flow of information between distinct network blocks, many studies [36], [37] have proposed cross-layer semantic interaction to integrate deep-layer global semantic information and shallow-layer local detail information. To adaptively integrate features from different network layers, several attention-based models have been proposed to distinguish the importance of different features and use the generated attentive features for salient object refinement [6], [38]. A

novel reverse attention network is devised to capture missing object information, improving the object integrity [39]. Wang et al. [40] proposed an approach to determine the salient object by analyzing fixation maps. To refine the coarse boundaries, researchers have proposed exploiting the boundary-enhancement loss function [41], [42] or learning to seamlessly integrate salient boundaries and multilevel salient regions [5], [43]. Inspired by human learning, Jin et al. [44] proposed a dual-stage self-paced strategy to enhance detection. The method first focuses on easy samples to establish a strong foundation and then gradually incorporates more challenging ones, adaptively weighting them on the basis of difficulty.

Although these deep CNN-based models have achieved impressive performance, their ability to obtain large receptive fields is still limited [45]. Many researchers have reported that both "Transformer" and graph convolution network (GCN) hold significant promise in overcoming the limited receptive field challenge. Many works have applied the Transformer backbone to acquire global context information by modeling long-range dependency [46], [47]. To integrate local information and global information, many works have proposed hybrid networks to exploite CNNs and Transformers [48]. In [49], the CNN and Transformer are grafted into a one-stage framework, and a novel cross-model attention fusion module guides the combination of complementary information more holistically. BCMNet [50] leverages a dual encoder with a CNN and Transformer to capture textures and contexts, facilitating concurrent feature fusion, morphology detection, and contour refinement via bidirectional collaborative mentoring. WaveNet [51] adopts a siamese encoder to extract features across modalities, subsequently fusing them via a discrete wavelet transform for both low- and high-frequency integration. FSANet [52] builds a dual-domain encoder to extract features from the frequency and spatial domains, attentively fusing them into hybrid features with enhanced discriminability. Many works [7], [53], [54] project the image into graph structure data, and devise the information aggregation network under the framework of GCNs, which are more flexible in modeling the long-range region relation. Yin and Lin [55] recently achieved efficient SOD by applying adder neural networks with a simple differential merging strategy. Although the traditional NSI-SOD models may fall short in addressing the distinct challenges posed by ORSIs, they remain a rich source of inspiration for researchers, fueling innovative approaches to advancing ORSI-SOD research.

B. Salient Object Detection in ORSIs

Owing to varying imaging conditions and environments, the salient objects in ORSIs exhibit complex geometric topologies, variable sizes, and cluttered backgrounds. Acknowledging the distinctive features of ORSIs, many researchers have put forward innovative insights and specifically designed numerous exceptional algorithms tailored for ORSI-SOD. Some works have conducted in-depth analysis on the scale variation of salient objects in ORSIs and explored several multiscale feature extraction and integration approaches, aiming to perceive salient objects of various scales [8], [9], [21]. Li et al. [8] designed a dual-stream pyramid structure to extract hierarchical and complementary information, enabling the perception of salient objects across diverse scales. Zhang et al. [9] devised a cascaded pyramid attention architecture, where shallow-layer attention and deep-layer attention cues can interact to perceive the salient object at different scales. Cong et al. [21] combined parallel multiscale attention in shallow-layers and spatial-wise and channel-wise relation reasoning in deep-layers, improving the integrity of salient objects across varying sizes. Li et al. [20] integrated multiple types of content through diverse attention techniques, leveraging the complementarity among various features such as, foreground, boundary, background, and global image features to enhance overall performance. Inspired by the edge-aided SOD in NSIs, many works deeply mine the edge/boundary-guided CNN architecture to refine the coarse boundary of the salient object with complicated edges [10], [19], [22]. Zhou et al. [22] devised an edge-guided recurrent positing framework equipped with two decoders: 1) one extracts edge information, and 2) the other locates the salient object by combining the edges at various scales. Zhou et al. [19] proposed incorporating features across multiple scales via guidance from edge cues. Tu et al. [10] developed a collaborative learning approach that incorporates bidirectional feature transformation, enabling the concurrent augmentation of both edge and region features. The adjacent cross-layer features are also explored to adaptively complement the local details and global information, enhancing the integrity of the salient regions [16]. To perceive the salient objects at a larger receptive field, Wang et al. [23] introduced a hybrid encoder by integrating CNNs and Transformer into a single-stream network, enabling the capture of the global context to mitigate the interference caused by complex backgrounds. Zhao et al. [56] devised an adaptive dual-stream encoder that addresses the issue of compensating for the global information obtained from Transformers and the local information obtained from CNNs. Huang et al. [24] enhanced object representation by modeling intrascene variations with cross-image context stored in foreground and background banks. Gu et al. [25] improved ORSI performance by enhancing features with bidirectional attention and mitigating attention loss via foreground-background decoupling. Liang and Luo [57] utilized a lightweight backbone network integrated with multiscale edge-embedded attention and multilevel semantic guidance to realize a lightweight ORSI-SOD model.

However, some challenges in ORSI-SOD, such as irregular geometric topological structures, scale variations within the same image, and cluttered backgrounds, remain unresolved by the aforementioned methods. To address these issues effectively, we introduce a novel CMNFNet. This framework leverages heterogeneous features to enhance performance in ORSI-SOD tasks, which is discussed in detail in the subsequent section.

III. Proposed Method

In this section, we delve deeper into our proposed CMNFNet. Section III-A introduces the overarching architecture of CMNFNet. Section III-B provides a concise overview

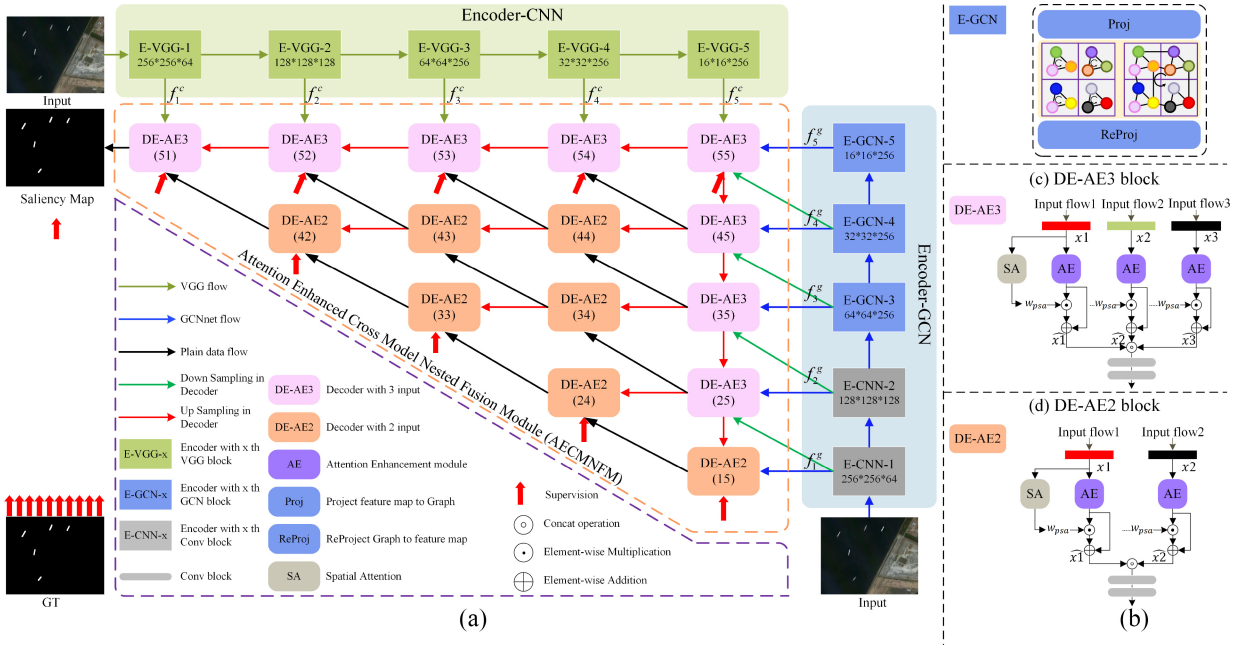

Fig. 2. Visual representations of the architecture. (a) Framework of our proposed CMNFNet. (b) Details of our E-GCN. (c) Details of our DE-AE3. (d) Details of our DE-AE2.

of the CNN-based feature extractor. Section III-C delves into the meticulously designed GCN-based feature extractor. Section III-D details the attention enhanced cross model fusion module. Section III-E describes the loss function.

A. Overall Architecture

As illustrated in Fig. 2(a), our CMNFNet is equipped with two heterogeneous encoder subnetworks (Encoder-CNN and Encoder-GCN) and an AECMNFM. The heterogeneous features (generated from our heterogeneous encoder) here actually represent distinct descriptions of the same input. By leveraging different information aggregation strategies, these encoders enhance feature diversity and complementarity. These heterogeneous features are subsequently fused in a nested fusion decoder, effectively addressing the challenges of SOD in ORSIs.

B. CNN-Based Feature Extractor

Similar to prior studies [9], [10], [16], [19], [22], we utilize the classical VGG-16 as the fundamental feature extractor. Unlike the primary VGG-16 architecture designed for image classification, we omit the final classification heads. The "Encoder-CNN" depicted at the top of Fig. 2(a) represents the CNN-based encoder subnetwork, which consists of five blocks denoted E-VGG-x ($x \in \{1, 2, 3, 4, 5\}$ is the block index). We utilize the outputs of the last convolution layer of these five blocks as side-output feature maps, denoted $f_i^c \in \mathbb{R}^{C_i \times H_i \times W_i}$ ($i \in \{1, 2, 3, 4, 5\}$). The values for $C_{1,2,3,4,5}$ are 64, 128, 256, 256, 256, and $H_{1,2,3,4,5}$ are 256, 128, 64, 32, 16. Note that, to alleviate computational complexity, we reduce the dimension of the side outputs in the final two blocks from 512 to 256 by introducing additional convolution layers.

C. GCN-Based Feature Extractor

To target salient objects with complex geometric topological structures and variable scales in remote sensing images, we design a novel context-aware GCN encoder, which exploits the GCN in modeling long-range dependencies between any two spatial regions. The "Encoder-GCN" shown in Fig. 2(a) provides an overview of our proposed GCN-based encoder subnetwork, which consists of two basic convolution blocks "E-CNN-1" and "E-CNN-2," and three specially devised graph convolution blocks "E-GCN-3," "E-GCN-4," and "E-GCN-5." The first two "E-CNN-x" ($x \in \{1, 2\}$ is the block index) blocks are all composed of Conv3×3, batch normalization (BN), and a ReLU, which are utilized to extract discriminative features and decrease the spatial resolution of the feature maps. Then, the feature maps are fed into three "E-GCN-x" blocks ($x \in \{3, 4, 5\}$ is the block index) sequentially, and the local–global context information is mined in graph embedding space. In each "E-GCN-x" block, shown in "GCN-based encoder block (E-GCN)" of Fig. 2(b), three operations are conducted sequentially. First, the feature maps are projected from regular grid data to irregular graph data. To mitigate the computational complexity associated with graph projection, we divide the feature maps into four patches, and project each patch into three graph nodes by graph projection operations. Considering the scale-variation objects within the same image, we construct local graph data and global graph data at different active regions, respectively. The local graph is built by linking every node within the same patch, whereas the global graph is built by linking all nodes across all patches. Second, to perceive scale-variation objects within the same image, graph reasoning operations are conducted on the local graph and global graph, and the parallel graph results are fused in an adaptive way. Finally, the acquired graph data are reprojected into regular

grid feature maps. The local–global context information for the salient objects with complex geometric topological structures and variable scales are modeled through three rounds of iterative operations. Similar to the CNN encoder, we also adopt the outputs of five blocks as side-out feature maps, which are denoted $f_i^g \in \mathbb{R}^{C_i \times H_i \times W_i} (i \in \{1, 2, 3, 4, 5\})$. $C_{1,2,3,4,5} = 64, 128, 256, 256, 256$ and $H_{1,2,3,4,5} = 256, 128, 64, 32, 16$.

1) Patch-Based Graph Projection: To construct graph structure data $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ from a regular grid feature map $f \in \mathbb{R}^{C \times H \times W}$, we propose learning a patch-based pixel-to-vertex soft assignment matrix, which clusters the pixels with similar features into coherent image regions. Different from the existing work [58] that produced a pixel-to-vertex soft assignment matrix by calculating the relationship between all pixels and all graph vertices, owing high computational complexity of $\boldsymbol{O}(\mathrm{HWC}|\mathcal{V}|)$. We propose grouping pixels into clusters within a smaller image patch individually. Specifically, given the regular grid feature map $f \in \mathbb{R}^{C \times H \times W}$, we first split it into $n \times n$ image patches. As depicted in the top part of Fig. 3, n is set to 2 by default in this work. In each image patch, we compute the similarity vector $r_{i,j}^{p,k}$ by measuring the Euclidean distance between pixel $f_{i,j}^p$ and the specific graph anchor point c_k^p. It is formulated in

$$r_{i,j}^{p,k} = f_{i,j}^p - c_k^p \tag{1}$$

where $f_{i,j}^p \in \mathbb{R}^{C \times 1}$ refers to the (i, j)th pixel in pth patch, and $c_k^p \in \mathbb{R}^{C \times 1}$ is the initial feature vector of the kth graph vertex in the pth patch. Then, we generate the pixel-to-vertex soft assignment value via a Softmax operation on all K graph vertices within an image patch. It is formulated in

$$q_{i,j}^{p,k} = \frac{\exp\left(-\|r_{i,j}^{p,k}/s_k^p\|_2^2\right)}{\sum_k \exp\left(-\|(r_{i,j}^{p,k})/s_k^p\|_2^2\right)} \tag{2}$$

where $s_k^p \in \mathbb{R}^{C \times 1}$ is a learnable column vector. $\|\cdot\|_2^2$ represents L2 normalization. We also represent $\mathcal{Q} \in \mathbb{R}^{hw \times |\mathcal{V}|}$ as the final complete soft assignment matrix from pixels to vertices, where h and w are the height and width of image patches, respectively, and $|\mathcal{V}|$ refers to the total number of graph vertices. In each row, the vector $q_{i,j}^{p,k}$ belonging to the same pth image patch satisfies that $\sum_k q_{i,j}^{p,k} = 1$. Finally, we encode feature v_k^p for the kth graph vertex in the pth image patch by weighted averaging of the residuals $r_{i,j}^{p,k}$. It is in

$$v_k^p = \frac{1}{\sum_{i,j \in p} q_{i,j}^{p,k}} \sum_{i,j \in p} q_{i,j}^{p,k} r_{i,j}^{p,k} / s_k^p. \tag{3}$$

We also denote $\mathcal{V} \in \mathbb{R}^{C \times |\mathcal{V}|}$ as the final obtained graph embedding feature, where $|\mathcal{V}| = n \times n \times K$ represents the total count of graph vertices and K is the number of graph vertices in each image patch. Computational complexity is reduced from $\boldsymbol{O}(\mathrm{HWC}|\mathcal{V}|)$ to $\boldsymbol{O}(\mathrm{HWC}|\mathcal{V}|/n^2)$.

2) Local–Global Graph Construction: To address the challenge caused by salient objects of varying scales within the same image, we propose constructing a local–global context-aware graph structure data at different active patch regions. Specifically, in each image patch, we generated K graph vertices by operating above patch-based graph projection G_{proj}. To better perceive the small salient objects (SSOs), we construct the local graph data $\mathcal{G}_L^r = (\mathcal{V}_L^r, \mathcal{E}_L^r) \in \mathbb{R}^{C \times K}$ $(r \in \{1, 2, 3, 4\})$ by connecting each graph vertex within the same image patch, which is shown in the left-top part of Fig. 3. The adjacent matrix $\mathrm{Adj}_L^r \in \mathbb{R}^{K \times K}$ can also be generated by

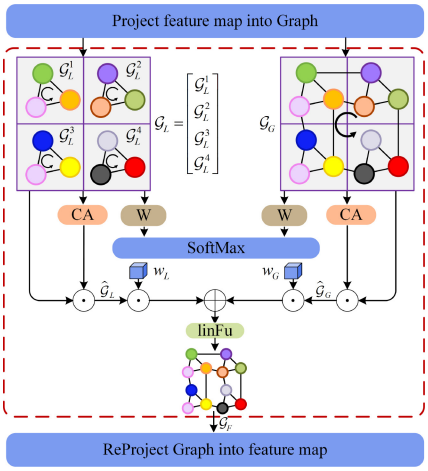

Fig. 3. Illustration of E-GCN. This module projects 2-D feature maps into local-connected and global-connected graph data, and conducts two parallel graph convolutions on them with different receptive fields. The graph features are adaptively fused by weight, and then reprojected into 2-D feature maps.

$$\mathrm{Adj}_L^r = \left(\mathcal{G}_L^r\right)^T \left(\mathcal{G}_L^r\right) \tag{4}$$

where $(\cdot)^T$ represents the matrix transpose operation. Similarly, to better perceive the large size salient object, we construct the global graph data $\mathcal{G}_G = (\mathcal{V}_G, \mathcal{E}_G) \in \mathbb{R}^{C \times |\mathcal{V}|}$ by connecting all graph vertices over all image patches, which is shown in the right-top part of Fig. 3. The adjacent matrix $Adj_G \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ can also be generated by

$$\mathrm{Adj}_G = (\mathcal{G}_G)^T \mathcal{G}_G. \tag{5}$$

It is noteworthy that both the parallel local and global graphs in Fig. 3 are built on the same set of graph nodes. The purpose is to map these nodes into two distinct graph spaces for information aggregation, thereby enabling the perception of both small and large targets. Subsequent ablation experiments in Section IV-C1 clearly show that the removal of either the local or global graph leads to a performance degradation.

3) Graph Reasoning: The local graph $\mathcal{G}_L$ or global graph $\mathcal{G}_G$ has a receptive field of all vertices on the image patch or on the whole image, respectively, it is potential for capturing local and global context information for scale-variation salient objects. Thus, we use the Edge-Gated Graph Convolution Operation [53] to further propagate important information on the local and global graph data adaptively. The information aggregation is formulated in

$$v_i^{l+1} = \sum_{j \in \mathcal{N}_i^l} eg_{i,j} \odot M^l v_j^l + U^l v_i^l \tag{6}$$

where v_i^{l+1} refers to the newly updated feature of the ith graph vertex in the $(l+1)$th layer, and v_i^l and v_j^l refer to the features of the ith and jth graph vertices in the lth layer. $\odot$ indicates the

element-wise Hadamard product. $eg_{i,j}$ is the Edge-Gated filter, which can promote the transmission of crucial information while suppressing the transmission of irrelevant information in graph data. $M^l \in \mathbb{R}^{C\times C}$ and $U^l \in \mathbb{R}^{C\times C}$ represent the learnable parameters. $\mathcal{N}_i^l$ represents the neighbor vertices set for the ith central graph vertex.

4) Adaptive Weighted Graph Fusion: The scale-variation salient objects have different sensitivities to local and global context information within the same image. To comprehensively model the local–global context information, we distinguish the importance of local graph embedding $\mathcal{G}_L$ and global graph embedding $\mathcal{G}_G$ for the salient objects of different scales and then generate the final graph embedding through adaptive weighted graph fusion. Specifically in Fig. 3, we first conduct channel-wise attention ($CA(\mathcal{G})$) on two graph structure data using (7), which enhances the graph representation $\hat{\mathcal{G}}$ by a residual connection in

$$\begin{cases} CA(\mathcal{G}) = \sigma(\text{MLP}(\text{global\_max\_pool}(\mathcal{G}))) & (7) \\ \hat{\mathcal{G}} = \mathcal{G}(1 + CA(\mathcal{G})) & (8) \end{cases}$$

where global_max_pool is the channel-wise max pooling operation along the graph nodes, which generates the channel-wise attention weight in $\mathbb{R}^{1\times C}$. The multilayer-perception (MLP) has parameters in $\mathbb{R}^{C\times C}$. σ is a sigmoid function.

To evaluate the importance of two graph embeddings, we conduct weight generation ($W(\mathcal{G})$) to produce the weight of each graph embedding, which is formulated in

$$W(\mathcal{G}) = \sigma(\text{linear}(\text{global\_mean\_pool}(\mathcal{G}))) \quad (9)$$

where global_mean_pool indicates the channel-wise mean pooling operation along the graph nodes, which generates the channel-wise attention weight in $\mathbb{R}^{1\times C}$. *linear* is the linear transformation with parameters in $\mathbb{R}^{C\times 1}$, which project the channel attention into a single weight. Then, we adopt the Softmax operation to balance the local graph embedding and global graph embedding, which are formulated as

$$\begin{cases} w_L = \dfrac{\exp(W(\mathcal{G}_L))}{\exp(W(\mathcal{G}_L)) + \exp(W(\mathcal{G}_G))} & (10) \\ w_G = \dfrac{\exp(W(\mathcal{G}_G))}{\exp(W(\mathcal{G}_L)) + \exp(W(\mathcal{G}_G))}. & (11) \end{cases}$$

Finally, the final graph embedding $\mathcal{G}_F$ is obtained by averaging the enhanced local graph embedding $\hat{\mathcal{G}}_L$ and the global graph embedding $\hat{\mathcal{G}}_G$ weighted by w_L and w_G, respectively. It is formulated in

$$\mathcal{G}_F = \text{linFu}\left(w_L \cdot \hat{\mathcal{G}}_L + w_G \cdot \hat{\mathcal{G}}_G\right) \quad (12)$$

where w_L and w_G refer to the importance of local graph and global graph, respectively. *linFu* is a linear function.

5) Graph Reprojection: To reproject the irregular graph data into 2-D grid feature map f, we conduct the graph reprojection operation G_{reproj} on each image patch. G_{reproj} can linearly interpolate 2-D pixel features on the basis of their region assignment $Q^r \in \mathbb{R}^{(hw)\times K}$. Specifically, given the rth image patch, we compute the pixel features of $\hat{f}^r \in \mathbb{R}^{C\times h\times w}$ by reweighting the graph vertices features $\mathcal{G}_F^r \in \mathbb{R}^{C\times K}$, formulated as

$$\hat{f}^r = \text{view}\left(\mathcal{G}_F^r (Q^r)^T\right) \quad (13)$$

where $r \in \{1, 2, 3, 4\}$ is the index of the image patch and $(\cdot)^T$ represents the matrix transpose operation. $\mathcal{G}_F^r \in \mathbb{R}^{C\times K}$ refers to the updated graph embedding belonging to the rth image patch. $Q^r \in \mathbb{R}^{(hw)\times K}$ refers to the soft assignment matrix belonging to the rth image patch. view reshapes it into a shape of $\mathbb{R}^{C\times h\times w}$.

D. Attention Enhanced Cross Model Nested Fusion Module

The heterogeneous features (extracted from the CNN-based encoder and GCN-based encoder) exhibit diversity and complementarity. However, traditional direct fusion (Dual-Stream or Single-Stream fusion strategy) of these features often leads to mutual interference and information loss, hindering their effective complementary integration. Our Nested Fusion is inspired by the notion that progressively fusing heterogeneous features at different levels can reduce mutual interference, minimize information loss, and facilitate their complementary fusion. Specifically, we propose encoding robust salient features by progressively selecting two kinds of heterogeneous features via nested fusion. The proposed AECMNFM is shown in Fig. 2, where attention-enhanced decoder submodule (DE-AE) is treated as the basic component and organized in a nested architecture. As shown in Fig. 2(c) and (d), the heterogeneous features are fed into "attention enhanced decoder block with three input (DE-AE3)" or "attention enhanced decoder block with two inputs (DE-AE2)," from which these features are first enhanced by the proposed "AE" and spatial attention "SA," then we concatenate the enhanced heterogeneous features and conduct convolution operation. Note that, the basic configurations of DE-AE3 and DE-AE2 are the same, and the sole difference between them is the number of inputs they can receive. The features follow the data flow shown in the "AECMNFM" of Fig. 2. Next, we first present the operations in "DE-AE" and then describe the key attention components in "AE."

1) Attention-Enhanced Decoder Submodule: Taking DE-AE3 in Fig. 2(c) as an example, for the given heterogeneous features $x1, x2, x3$, the plain spatial attention (PSA) weight w_{psa} is first generated by conducting PSA F_{psa} (formulated in (14)) on feature $x1$. Then, the enhanced features $\hat{x1}, \hat{x2}, \hat{x3}$ are enhanced by the attention enhanced ("AE") module and spatial attention w_{psa}, using (15)-(17). Finally, the enhanced features are concatenated and fused by performing two basic convolution operations, which are formulated in

$$\begin{cases} w_{\text{psa}} = F_{\text{psa}}(x1) & (14) \\ \hat{x1} = AE(x1)(1 + w_{\text{psa}}) & (15) \\ \hat{x2} = AE(x2)(1 + w_{\text{psa}}) & (16) \\ \hat{x3} = AE(x3)(1 + w_{\text{psa}}) & (17) \\ x_{de} = \text{conv}(\text{concat}(\hat{x1}, \hat{x2}, \hat{x3}) & (18) \end{cases}$$

The F_{psa} refers to the PSA operation described in (25). Concat concatenates the features in channel dimension. conv indicates the convolution layer. The AE represents the proposed attention enhanced module (AE-M), which is shown in Fig. 4 (a) and can be formulated as

$$AE = \text{PSA}(\text{HCA}(x)) \quad (19)$$

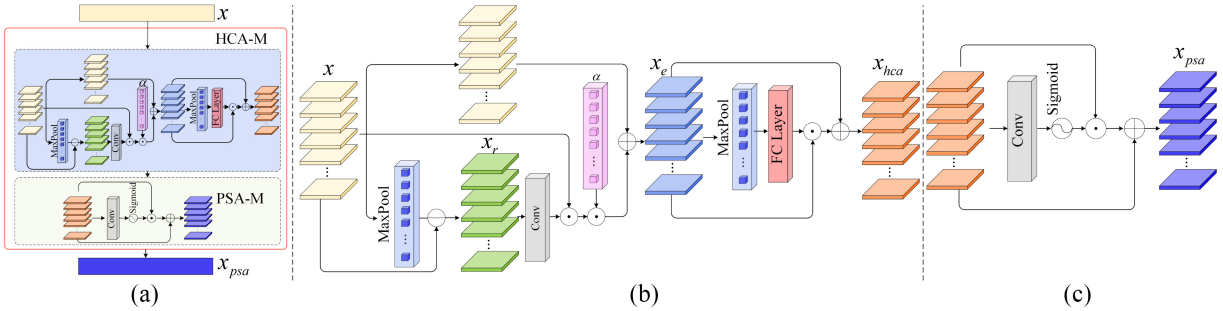

Fig. 4. Illustration of our AE-M. It is composed of HCA, which explores the plain and reverse attention to enhance channel-wise features, and PSA, which further enhances spatial-wise features. (a) AE-M. (b) HCA-M. (C) PSA-M.

where HCA and PSA refer to the following hybrid channel attention (HCA) submodule and PSA (PSA) submodule, respectively.

2) Hybrid Channel Attention: By inspecting the feature maps channel by channel, we found that some effective feature maps can accentuate the foreground and attenuate the background, whereas some adverse feature maps incorrectly accentuate the background and attenuate the foreground. Directly applying traditional channel attention may still suffer from the cluttered background, so we devise a HCA module to enhance the representations by weakening the contrary feature maps. Specifically, we keep the effective feature maps unchanged, while reversing the contrary feature maps to explore effective information as much as possible. As shown in Fig. 4(b), to generate reversed feature maps, we first compute the channel-wise maximum using (20), and obtain the reversed feature maps via

$$\begin{cases} \max = \max\_\text{Pool}(x) & (20) \\ x_r = \max - x & (21) \end{cases}$$

where $x \in \mathbb{R}^{C\times H\times W}$ refers to the feature maps. max_pool is the max pooling operation along the spatial dimension. $\max \in \mathbb{R}^{C\times 1\times 1}$ refers to the maximum value within each channel. $x_r \in \mathbb{R}^{C\times H\times W}$ represents the reversed feature maps. To obtain enhanced feature maps, we apply a weighted average of the original and reversed feature maps using

$$x_e = x + \alpha \odot (\text{conv}(x_r) \odot x) \quad (22)$$

where $\alpha \in \mathbb{R}^{C\times 1\times 1}$ is the learnable parameters, and each value indicates the importance of each channel. $x_e \in \mathbb{R}^{C\times H\times W}$ represents the enhanced feature maps. Then, the new channel-wise attention weights are generated using

$$w_{\text{hca}} = \text{FCL}(\max\_\text{Pool}(x_e)) \quad (23)$$

where FCL is the fully connected layer. The new feature x_{hca} enhanced by channel attention can be obtained by

$$x_{\text{hca}} = x_e + w_{\text{hca}} \odot x_e \quad (24)$$

where $\odot$ refers to the channel-wise Hadamard product.

3) Plain Spatial Attention: In order to focus on salient objects in the spatial dimension, as shown in Fig. 4(c), we generate a spatial attention map by performing convolution operations on x_{hca}. It is formulated in

$$F_{\text{psa}} = \sigma(\text{conv}(x_{\text{hca}})) \quad (25)$$

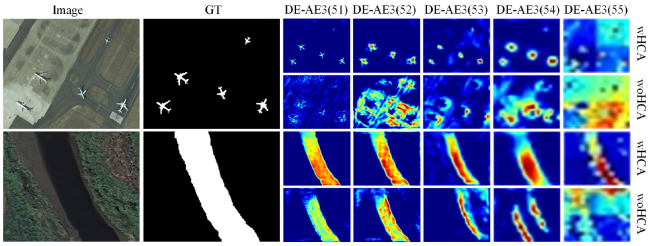

Fig. 5. Visual examples of the feature maps produced by different decoder blocks. The "wHCA" in the 1st and 3rd rows show the feature maps produced by the blocks equipped with HCA, and the "woHCA" in the 2nd and 4th rows show the feature maps produced by the blocks equipped without HCA.

where conv refers to the basic convolution operation, and σ is the sigmoid function. Then, the final feature maps enhanced by spatial attention can be obtained using

$$x_{\text{psa}} = x_{\text{hca}} + F_{\text{psa}}(x_{\text{hca}}) \odot x_{\text{hca}} \quad (26)$$

where $\odot$ is the Hadamard product over spatial dimension.

Fig. 5 presents the feature maps within five decoder blocks. This shows that with HCA in AE-M, the clutter background can be effectively suppressed and the salient objects can be more precisely focused.

E. Loss Function

To train our network more effectively, we utilize a combination of binary cross-entropy (BCE) loss and intersection over union (IOU) loss, which comprehensively supervises all nine saliency output of the network decoder, formulated as follows:

$$L = \sum_{i=1}^{9} (l_{\text{BCE}}(S_i, G) + l_{\text{IOU}}(S_i, G)) \quad (27)$$

where S_i is the ith saliency outcome produced by the ith decoder block, G is the ground truth maps corresponding to the original input images, and L represents the whole loss function.

IV. Experiment

A. Experimental Protocol

1) Datasets: To perform the following comparison experiments, we select three benchmark datasets.

ORSSD [8], the earliest dataset released for ORSI-SOD, comprises an assemblage of images sourced from Google

Earth and other existing ORSI datasets. It splits the dataset into training and testing sets, with sizes of 600 images and 200 images, respectively. Each image is accompanied by pixel-wise annotations for salient objects.

EORSSD [9] is an expanded version of ORSSD with 1400 images for training and 600 images for testing. It encompasses more complex and diverse scenes, with each image accompanied by pixel-wise annotations for salient objects.

ORSI4199 [10] is a recently released ORSI-SOD dataset with 2000 training images and 2199 testing images, all with precise pixel-level annotations. This dataset encompasses diverse complex objects across various attributes, rendering it one of the most challenging datasets.

2) Evaluation Metrics: Five mainstream performance metrics are utilized to assess our proposed network thoroughly.

S-Measure (S_α) [59] assesses the likeness between the saliency map and ground truth, including both object-specific (S_{oj}) and region-specific (S_{re}) structure similarity

$$S_\alpha = \alpha \times S_{oj} + (1-\alpha) \times S_{re} \tag{28}$$

where the balancing weight α is set to 0.5 by default.

F-Measure (F_β) [60] provides a balanced evaluation metric that considers Precision (Pr) and Recall (Re)

$$F_\beta = \frac{\left(1+\beta^2\right) \times Pr \times Re}{\beta^2 \times Pr + Re} \tag{29}$$

where we adopt $\beta = 0.3$ to align with mainstream standards.

E-measure (E_ξ) [61] measures both the pixel-level correspondence and image-level statistical information

$$E_\xi = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} \phi(x, y) \tag{30}$$

where H and W are the height and width of the map, respectively. $\phi(x, y)$ represents the enhanced alignment function.

MAE (M) [62] calculates the average deviation of all saliency maps from the corresponding ground truths

$$\text{MAE} = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} |S(x, y) - G(x, y)| \tag{31}$$

where $S(x, y)$ and $G(x, y)$ refer to the pixel values indexed by (x, y) in the saliency map and the ground truth, respectively. $|\cdot|$ represents the absolute value function.

PR Curve [60] depicts the relationship between precision (Pr) and recall (Re) across various thresholds. A curve that is closer to the top right corner indicates superior performance

$$Pr = \frac{S_b \bigcap G}{S_b}, Re = \frac{S_b \bigcap G}{G} \tag{32}$$

where S_b is the binarized salient segmentation map with different thresholds, and G is the ground truth map.

3) Implementation Details: We performed all the experiments on our workstation, which runs the Windows 10 operating system and is equipped with an NVIDIA GeForce RTX 3090 GPU, Python 3.8, and PyTorch 1.11.0+cu115. During the training stage, we resized both the input image and groundtruth to 256 × 256. We augmented the dataset eight times by flipping and rotating operations, as recommended by [8]. Additionally, the batch size was 8, and the model was trained for 100 epochs on each of the three training sets separately. We utilized the Adam optimizer with an initial learning rate of 0.0002 and halved it every 20 epochs for efficient training. In addition, we developed two versions of the model: 1) Ours-V, which employs VGG-16 as the CNN feature encoder, and 2) Ours-R, which uses ResNet-50 for the same purpose.

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

1) Models for Comparative Studies: Our proposed CMNFNet and 16 state-of-the-art (SOTA) models are compared across all three benchmark datasets. The compared methods encompassed a diverse range of models: RRWR [30] and RCRR [28] are two conventional NSI-SOD models. EG [5], MI [4], and Gated [18] are three deep NSI-SOD models. LV [8], DAF [9], MJRB [10], EMFI [19], Corr [17], ERP [22], ACCo [16], MEA [57], and ADST [56] are nine deep ORSI-SOD models. RR [21] and HFA [23] are two Hybrid ORSI-SOD models. Table I lists all the quantitative results, which were generated by running the corresponding open-source codes provided by the author and adopting the default parameter configurations, or through calculations based on publicly accessible saliency maps. Notably, the results of the LV are unavailable for the ORSI4199 dataset, which is indicated as "-" in Table I. Our model demonstrates impressive speed, reaching 19 fps when paired with the VGG backbone and 18 fps when coupled with the ResNet backbone.

2) Quantitative Comparisons and Discussions: As shown in Table I, our approach (both Ours-V and Ours-R) generally achieves the nine best performances and three second best performances among all twelve evaluation metrics, which clearly demonstrates the superiority of our proposed method. Specifically, both Ours-V and Ours-R outperform all other models across all performance metrics on the ORSSD dataset. This validates the optimality of our method on this dataset. For the EORSSD dataset, although Ours-V achieves a slightly lower performance than DAF-V does with respect to E_ξ and MAE, Ours-V achieved 2.10% and 2.45% performance improvement with respect to S_α and F_β, respectively. Similarly, Ours-R ranks first with respect to E_ξ and MAE and rank second with regard to S_α and F_β. HFA-R achieves slightly higher performance than Ours-R dose with respect to S_α and F_β, which may be attributed to its larger input image size of 480 × 480.

For the ORSI4199 dataset, both Ours-V and Ours-R rank first with respect to F_β, E_ξ, and MAE. Although ACCoNet-V and ACCoNet-R outperform our method with respect to S_α, the gap is not significant. Notably, compared with the two most recent advanced models (MEA-MV2 [57] and ADST-R2 [56]), our method (Ours-R and Ours-V) also achieves superior performance across all datasets, which verifies the advantages of our method.

We also compare these methods with respect to the PR curves on three benchmark datasets, as shown in Fig. 6. Ours-V and Ours-R are indicated by bold red solid and dashed lines, respectively. The results presented in Fig. 6 show that

TABLE I

QUANTITATIVE COMPARISON BETWEEN OUR CMNFNET AND 16 SOTA MODELS ACROSS THREE BENCHMARK DATASETS. THE COMPARED MODELS FALL INTO FOUR CATEGORIES: TRADITIONAL NSI-SOD (T-NSI), DEEP NSI-SOD (D-NSI), DEEP ORSI-SOD (D-ORSI), AND HYBRID ORSI-SOD (H-ORSI) METHODS. THE MODELS APPENDED WITH "-V," "-R," "-R2," OR "-MV2" SUFFIXES SIGNIFY THE USE OF VGG, RESNET, RES2NET OR MOBILENETV2 AS THE CNN-BASED ENCODER, RESPECTIVELY. ↑ AND ↓ INDICATE THAT HIGHER AND LOWER VALUES ARE BETTER. THE TOP THREE PERFORMANCES ARE MARKED IN RED, BLUE, AND GREEN

Methods		RRWR [30]	RCRR [28]	EG [5]	MI [4]	Gated [18]	LV-V [8]	DAF-V [9]	MJRB-V [10]	EMFI-V [19]	Corr-V [17]	ERP-V [22]	ACCo-V [16]	**Ours-V**	MJRB-R [10]	EMFI-R [22]	RR-R [21]	HFA-R [23]	ERP-R [22]	ACCo-R [16]	MEA-MV2 [57]	ADST-R2 [56]	**Ours-R**
Year		2015	2018	2019	2020	2020	2019	2021	2022	2022	2022	2023	2023	-	2022	2022	2022	2022	2023	2023	2024	2024	-
Type		T-NSI	T-NSI	D-NSI	D-NSI	D-NSI	D-ORSI	D-ORSI	D-ORSI	D-ORSI	D-ORSI	D-ORSI	D-ORSI	H-ORSI	D-ORSI	D-ORSI	H-ORSI	H-ORSI	D-ORSI	D-ORSI	D-ORSI	D-ORSI	H-ORSI
FPS↑		0.3	0.3	-	12	25	1.4	26	32	25	100	50	81	19	22	25	109	26	50	-	23.75	39.5	18
ORSSD	S_α ↑	0.6835	0.6849	0.8721	0.9040	0.9186	0.8815	0.9191	0.9204	0.9366	0.9380	0.9254	0.9437	**0.9498**	0.9211	0.9432	0.9339	0.9399	0.9352	0.9428	0.9340	0.9379	**0.9475**
	F_β ↑	0.5590	0.5591	0.8332	0.8761	0.8871	0.8263	0.8928	0.8842	0.9002	0.9129	0.8974	0.9149	**0.9221**	0.8885	0.9155	0.9011	0.9117	0.9036	0.9149	0.9042	0.9124	**0.9189**
	E_ξ ↑	0.7649	0.7651	0.9731	0.9545	0.9664	0.9456	0.9771	0.9623	0.9737	0.9790	0.9710	0.9796	**0.9827**	0.9686	0.9813	0.9722	0.9770	0.9738	0.9819	0.9717	0.9807	**0.9832**
	M ↓	0.1324	0.1277	0.0216	0.0144	0.0137	0.0207	0.0113	0.0163	0.0109	0.0098	0.0135	0.0088	**0.0077**	0.0145	0.0095	0.0113	0.0092	0.0114	0.0087	0.0098	0.0086	**0.0078**
EORSSD	S_α ↑	0.5992	0.6007	0.8601	0.9040	0.9114	0.8630	0.9166	0.9197	0.9290	0.9289	0.9210	0.9290	**0.9376**	0.9091	0.9319	0.9266	**0.9380**	0.9252	0.9302	0.9282	0.9311	0.9377
	F_β ↑	0.3993	0.3995	0.7880	0.8344	0.8566	0.7794	0.8614	0.8656	0.8720	0.8778	0.8632	0.8837	**0.8859**	0.8555	0.8742	0.8743	**0.8876**	0.8743	0.8821	0.8844	0.8804	0.8851
	E_ξ ↑	0.6894	0.6882	0.9570	0.9442	0.9610	0.9254	**0.9861**	0.9646	0.9711	0.9696	0.9603	0.9727	0.9755	0.9655	0.9712	0.9665	0.9740	0.9665	0.9759	0.9717	0.9769	**0.9774**
	M ↓	0.1677	0.1644	0.0110	0.0093	0.0095	0.0146	**0.0060**	0.0099	0.0084	0.0083	0.0089	0.0074	0.0069	0.0099	0.0075	0.0082	0.0071	0.0082	0.0067	0.0070	0.0065	**0.0063**
ORSI4199	S_α ↑	0.6416	0.6490	0.8516	0.8232	0.8660	-	0.8477	0.8593	0.8688	0.8626	0.8642	**0.8775**	0.8766	0.8582	0.8712	0.8585	0.8767	0.8617	**0.8805**	0.8677	0.8710	0.8774
	F_β ↑	0.5405	0.5479	0.8371	0.7891	0.8443	-	0.8169	0.8493	0.8564	0.8560	0.8551	0.8687	**0.8734**	0.8511	0.8636	0.8500	0.8700	0.8500	0.8688	0.8591	0.8698	**0.8752**
	E_ξ ↑	0.7115	0.7192	0.9241	0.8961	0.9256	-	0.9201	0.9311	0.9338	0.9333	0.9284	0.9412	**0.9476**	0.9343	0.9403	0.9286	0.9431	0.9252	0.9424	0.9375	0.9433	**0.9485**
	M ↓	0.1718	0.1638	0.0385	0.0473	0.0387	-	0.0473	0.0374	0.0334	0.0366	0.0376	0.0314	**0.0305**	0.0372	0.0313	0.0367	0.0314	0.0405	0.0320	0.0333	0.0318	**0.0301**

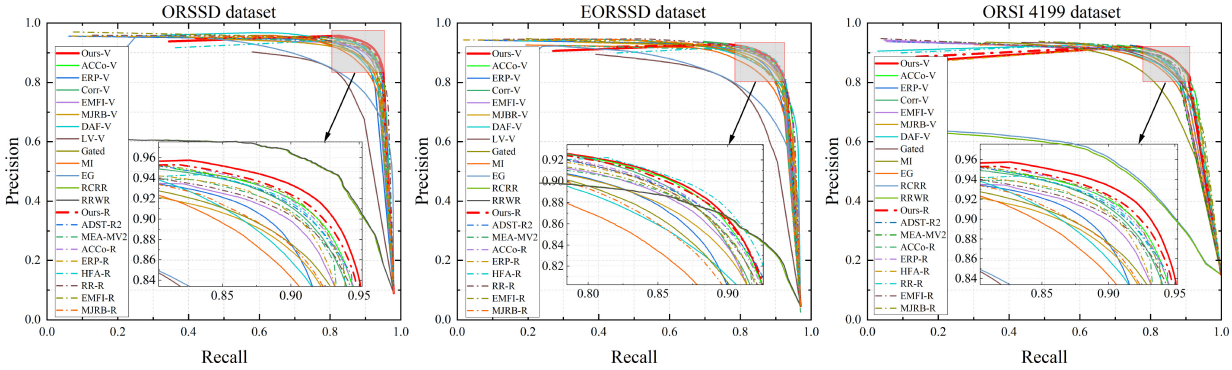

Fig. 6. Comparative analysis of the PR curves of our CMNFNet and 16 SOTA methods across three benchmark datasets.

our approach exhibits the closest proximity to the top right corner on three datasets, which further shows its superiority. Overall, both Ours-V and Ours-R generally achieve better performance than 16 SOTA models on the three datasets, which demonstrates the superiority of our CMNFNet.

3) Qualitative Comparisons and Discussions: We also qualitatively compare visual examples selected from the ORSSD, EORSSD, and ORSI4199 datasets. As shown in Fig. 7, all the visual examples are carefully chosen to encompass a diverse range of complicated scenes, such as background interference (e.g., the examples of ORSSD-1 and EORSSD-2), multiple salient objects (MSOs) of different sizes within the same image (e.g., the examples of ORSSD-3, EORSSD-1, and EORSSD-3), and complicated topological structures (e.g., the examples of ORSSD-2, ORSI4199-1, ORSI4199-2, and ORSI4199-3). Note that, the example in the *i*th row of each dataset is abbreviated as "data-i" in the following article.

For the examples of background interference shown in ORSSD-1 and EORSSD-2, the salient objects and background may share similar shapes or color features. Previous advanced methods may suffer from these cluttered backgrounds, and may mistakenly detect the background region near the salient object. Our method alleviates this issue through the AE-M in the cross-model nested fusion subnetwork, which suppresses background and complements heterogeneous features by selecting important salient features.

For the examples of MSOs of different sizes within the same image shown in ORSSD-3, EORSSD-1, and EORSSD-3, previous advanced methods may struggle to locate all the salient objects simultaneously, as seen in the cases of ACCo-V, ERP-V, and Corr-V. Conversely, our approach can discover all the salient objects with different scales and predict precise saliency maps on these challenging visual scenes. This is attributed to our tailored GCN-based encoder, which captures multiscale salient objects across diverse regions via dual parallel graph inference branches.

For the example of salient objects with complex appearance topological structures. ORSSD-2 and ORSI4199-3 exhibit salient objects with elongated shapes, which span the entire image and make it difficult for conventional detection methods to detect the entire salient object, resulting in partially missing predictions. ORSI4199-1 and ORSI4199-2 exhibit salient objects with complex topological shapes, which makes it difficult for previous methods to capture complete topological structures, leading to misjudgments and some nonsalient areas being incorrectly highlighted. However, our proposed method can successfully handle these challenging visual scenes. This can be attributed to our specially devised two parallel graph inference branches in the GCN-based encoder, where the local

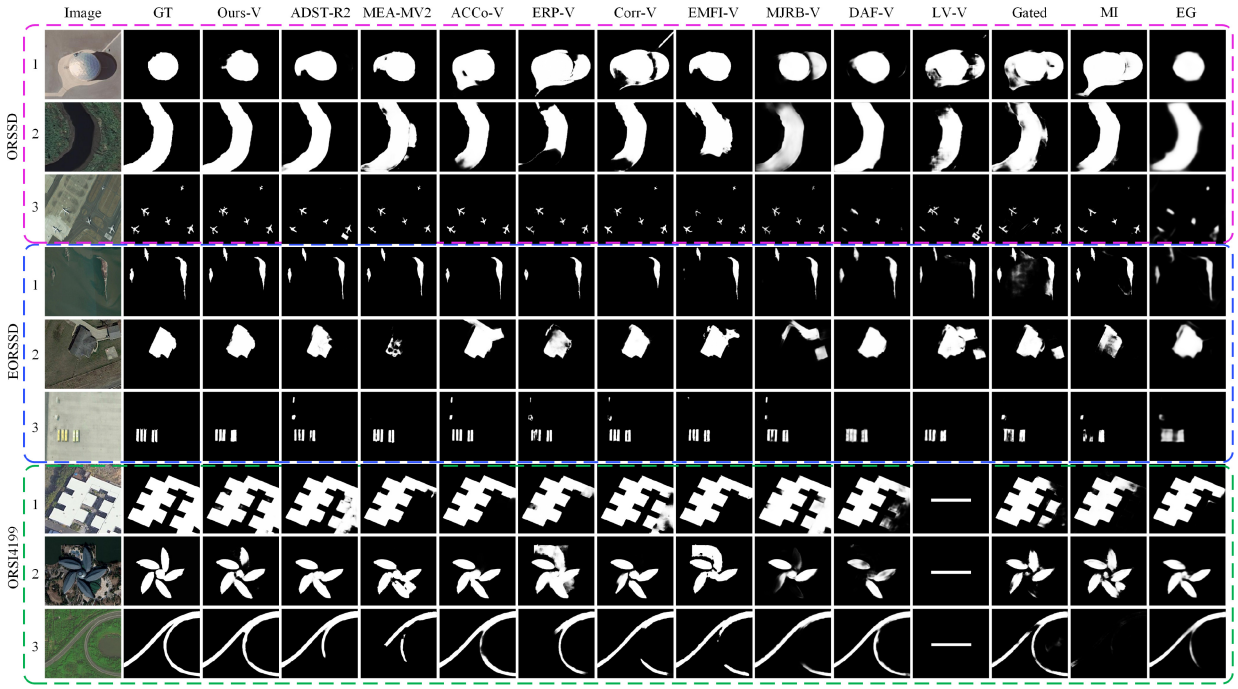

Fig. 7. Visual comparison between Ours method and other SOTA models across three diverse and challenging datasets. The suffix (-V, MV2, R2) appended to the method's name indicates that the model employs VGG or MobileNet-V2 or Res2Net as its CNN-based encoder. The visual examples presented encompass a diverse range of complicated scene, including those with background interference (ORSSD-1, EORSSD-2), MSOs of varying sizes within a single image (ORSSD-3, EORSSD-1, and EORSSD-3), and complicated topological structures (ORSSD-2, ORSI4199-1, ORSI4199-2, and ORSI4199-3).

TABLE II
SCENE SUBATTRIBUTE ANALYSIS ON THE ORSI4199 DATASET. TO ENSURE CONSISTENCY WITH ADVANCED METHODS, ALL THE MODELS USE THE RESNET BACKBONE, EXCEPT FOR THE CORR (VGG-ONLY), MEA (MOBILENET-V2-ONLY), AND ADST (RES2NET-ONLY) MODELS. THE TOP THREE MAX F_β VALUES ARE HIGHLIGHTED IN RED, BLUE, AND GREEN, RESPECTIVELY

Methods	Scene sub-attributes									
	BSO	CS	CSO	ISO	LCS	MSO	NSO	OC	SSO	Avg
MJRB [10]	0.9166	0.8961	0.8899	0.8937	0.7916	0.8585	0.8748	0.8422	0.8091	0.8636
RR [21]	0.8811	0.8702	0.8649	0.8594	0.7710	0.8288	0.8392	0.8042	0.7682	0.8319
Corr [17]	0.8891	0.8835	0.8728	0.8742	0.7852	0.8528	0.8724	0.8409	0.8051	0.8529
HFA [23]	0.9103	0.8964	**0.8948**	0.9046	0.8030	0.8566	0.8858	**0.8506**	0.8128	0.8682
ERP [22]	0.8946	0.8800	0.8727	0.8837	0.7925	0.8468	0.8467	0.8378	0.8015	0.8507
ACCo [16]	0.9091	0.8974	0.8931	0.8997	0.8048	0.8611	0.8814	0.8337	0.8065	0.8652
MEA [57]	0.9068	0.8911	0.8763	0.8801	0.7879	0.8533	0.8577	0.8395	0.8076	0.8556
ADST [56]	0.9180	0.9012	0.8845	0.8993	0.8092	0.8567	0.8895	0.8403	0.8055	0.8671
Ours	**0.9235**	**0.9059**	0.8940	**0.9097**	**0.8093**	**0.8618**	**0.9018**	0.8458	**0.8133**	**0.8739**

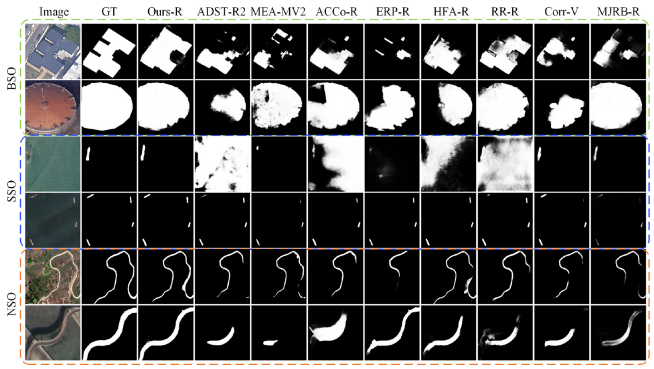

Fig. 8. Visual examples of the attribute analyzes on the ORSI-4199 dataset, compared with current advanced methods. The suffix (-R, V, MV2, R2) indicates that the model employs ResNet or VGG or MobileNet-V2 or Res2Net as its CNN-based encoder. The presented classic complex scenarios encompass BSO, SSO, and NSO scenes.

graph inference branch can explore the structural information of different parts of the salient object in different active regions, and the global graph inference branch can comprehensively grasp the overall structural information of the salient object in a global receptive field. The combined local structure details and global structure information make it more precise for predicting salient objects with complex topological structures.

4) Scene Subattribute Analysis: The ORSI-4199 dataset serves as a formidable benchmark, featuring a thorough categorization into various attribute subsets, which are carefully crafted to reflect the wide range of challenges presented in remote sensing scenes. We also compare our method with eight cutting-edge approaches across these diverse attribute challenges. All the max F_β results are presented in Table II, showing that our method achieves the seven best performances and 2 s best performances among the nine challenging attributes, and performs best in terms of average performance. These quantitative results show that our approach generally outperforms existing advanced models in scenes with different category attributes, highlighting its effectiveness in handling various challenges.

Specifically, our method significantly outperforms other methods in scenarios featuring big salient objects (BSOs) and narrow salient objects (NSOs), highlighting its effectiveness in capturing complex topological structures. The BSO and NSO shown in Fig. 8 also validate the effectiveness of our approach in preserving the overall integrity of the salient object. In

TABLE III
ABLATION STUDIES CONDUCTED ON THE ORSSD AND EORSSD DATASETS. ↑ AND ↓ INDICATE THAT HIGHER AND LOWER VALUES ARE BETTER, RESPECTIVELY. SCORES HIGHLIGHTED IN BOLD INDICATE THE BEST, WHILE SCORES HIGHLIGHTED IN RED REPRESENT PERFORMANCE DEGRADATION

No.	E-GCN		AE-M			FusionStrategy			ORSSD						EORSSD					
	$\mathcal{G}_L$	$\mathcal{G}_G$	HCA	PCA	SA	Nest	2-S	1-S	$S_\alpha\uparrow$	$\triangle S_\alpha$	$F_\beta\uparrow$	$\triangle F_\beta$	$E_\xi\uparrow$	$\triangle E_\xi$	$S_\alpha\uparrow$	$\triangle S_\alpha$	$F_\beta\uparrow$	$\triangle F_\beta$	$E_\xi\uparrow$	$\triangle E_\xi$
1	✓	✓	✓		✓	✓			**0.9498**	**0.0000**	**0.9221**	**0.0000**	**0.9827**	**0.0000**	**0.9376**	**0.0000**	**0.8859**	**0.0000**	**0.9755**	**0.0000**
2									0.9230	-0.0268	0.8876	-0.0345	0.9636	-0.0191	0.9246	-0.0130	0.8618	-0.0241	0.9558	-0.0197
3	✓		✓		✓	✓			0.9451	-0.0047	0.9171	-0.0050	0.9799	-0.0028	0.9366	-0.0001	0.8852	-0.0007	0.9741	-0.0014
4		✓	✓		✓	✓			0.9449	-0.0049	0.9188	-0.0033	0.9809	-0.0018	0.9348	-0.0028	0.8829	-0.0030	0.9709	-0.0046
5	✓	✓			✓	✓			0.9406	-0.0092	0.9104	-0.0117	0.9777	-0.0050	0.9291	-0.0085	0.8710	-0.0149	0.9637	-0.0118
6	✓	✓	✓			✓			0.9460	-0.0038	0.9168	-0.0053	0.9802	-0.0025	0.9354	-0.0022	0.8843	-0.0016	0.9721	-0.0034
7	✓	✓		✓	✓	✓			0.9420	-0.0078	0.9152	-0.0069	0.9812	-0.0015	0.9345	-0.0031	0.8813	-0.0046	0.9701	-0.0054
8	✓	✓	✓		✓		✓		0.9398	-0.0100	0.9089	-0.0132	0.9777	-0.0050	0.9336	-0.0040	0.8822	-0.0037	0.9729	-0.0026
9	✓	✓	✓		✓			✓	0.9437	-0.0061	0.9154	-0.0067	0.9803	-0.0024	0.9353	-0.0023	0.8840	-0.0019	0.9735	-0.0020

complex scenes (CSs), our approach also exhibits performance enhancements relative to the second-best model, proving its capacity to discover the salient object and mitigate background interference. In scenarios with MSOs and SSOs, our model outperforms all other models, and the examples in Fig. 8 also show consistent results. In scenarios involving incomplete salient objects (ISOs) and low-contrast scenes (LCSs), our approach performs better than do all other advanced models. Although the performance of our model in CSO and off center (OC) scenarios is not yet the best, the gap between our method and the top-performing method is rather insignificant.

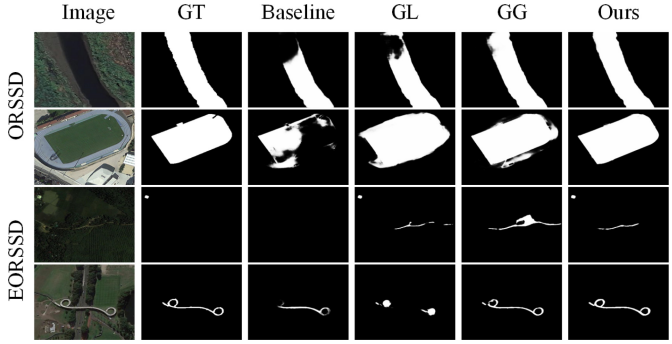

Fig. 9. Visual examples of the ablation study of local–global graph branches in the E-GCN block. "Baseline" represents the baseline model, "GL" represents the model equipped with only a local graph branch in the E-GCN, "GG" represents the model equipped with only a global graph branch in the E-GCN, and "Ours" represents the complete model equipped with both local and global graph branches in the E-GCN.

C. Ablation Study

This section presents ablation studies on the ORSSD and EORSSD datasets to evaluate the impact of our proposed modules with a VGG-based backbone.

1) Effect of the Local–Global Graph: The impact of the local and global graph reasoning branches in the E-GCN block is first assessed. Specifically, we trained two network variants equipped with only a local graph reasoning branch ($\mathcal{G}_L$) or only a global graph reasoning branch ($\mathcal{G}_G$). The quantitative results are shown in lines No.3 and No.4 of Table III. Compared to that of the complete network equipped with both local and global graph reasoning branches (shown in line No.1 of Table III), the performance of these two network variants consistently decreases with respect to all three metrics on both the ORSSD and EORSSD datasets. This finding validates the positive effect of our proposed two parallel graph inference branches on performance enhancement. We also present some examples generated by the baseline network variant, our complete network "Ours," the "$\mathcal{G}_L$" network variant, and the "$\mathcal{G}_G$" network variant. The examples presented in Fig. 9 show that "$\mathcal{G}_L$" still suffers from severely ISOs, whereas "$\mathcal{G}_G$" can maintain the integrity of the salient object and overlook certain local regions of salient objects. Our complete network 'Ours' equipped with both local and global graph branches can preserve the fine details and overall structure of the salient objects.

2) Effectiveness of the Attention Enhanced Decoder Block: We perform experiments to assess the influence of the attention modules (HCA and SA) in our attention enhanced decoder (DE-AE) block on enhancing performance. Specifically, we trained another three network variants in which "SA(woHCA)" representing HCA is removed from DE-AE, "HCA(woSA)" representing SA is removed from DE-AE, "PCA+SA" representing HCA is replaced with plain channel attention (PCA) in DE-AE, and "Ours" representing the complete model. The quantitative results are displayed in lines No.5, No.6, and No.7 of Table III. Compared with the model equipped with complete DE-AE (displayed in line No.1 Table III), all three metrics of these three network variants consistently decrease on both datasets. When we remove the HCA module from DE-AE (as shown in line No.5 of Table III), the performance significantly degrades in terms of both the F_β and E_ξ metrics. This validates the effectiveness of the HCA module in suppressing the background. When we remove the SA module from DE-AE (as shown in line No.6 of Table III), we observe a slight performance degradation in terms of all the metrics, confirming its role in highlighting important information. Replacing HCA with a PCA module (line No.7 in Table III) leads to performance degradation, highlighting the superior effectiveness of HCA in capturing channel-wise information.

Fig. 10 displays several examples. The visualization results show that regardless of whether the HCA module or the SA module is removed from the DE-AE, or if the HCA module is replaced with the PCA module, there are detection errors in the final predicted saliency maps. When we replace HCA with PCA, excessive salient object prediction occurs. In contrast,

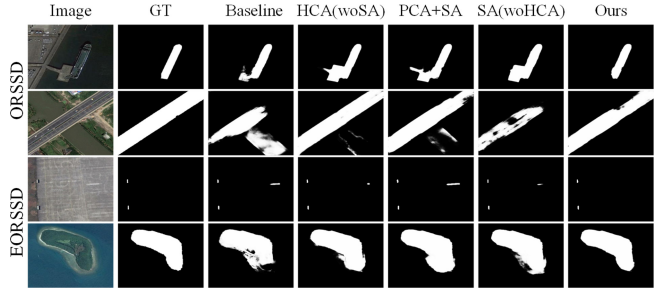

Fig. 10. Visual examples of the ablation study on DE-AE. "Baseline" represents the baseline model, "HCA(woSA)" represents that SA is removed from DE-AE, "PCA+SA" represents that HCA is replaced with PCA in DE-AE, "SA(woHCA)" represents that HCA is removed from DE-AE and "Ours" represents the complete model.

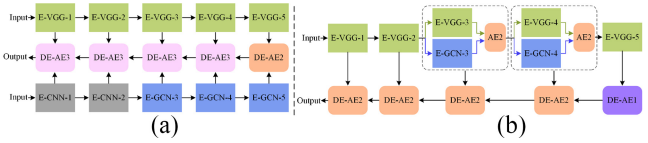

Fig. 11. Two traditional fusion strategies: Dual-Stream ("2-S," left) and Single-Stream ("1-S," right). (a) Dual stream fusion network (Abbreviated as symbol "2-S"). (b) Single-stream fusion network (Abbreviated as symbol "1-S").

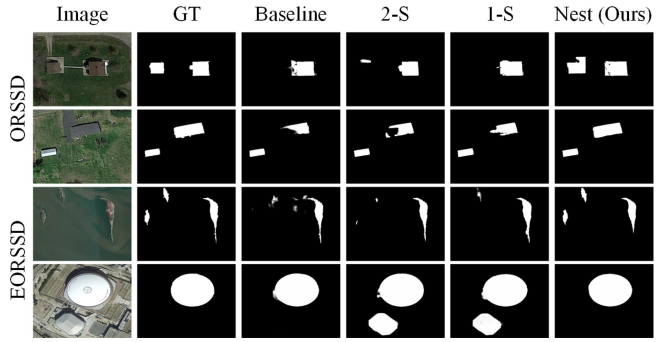

Fig. 12. Visual examples of ablation study of the fusion strategy. "Baseline" represents the baseline model, "2-S" represents the Dual-Stream strategy, and "1-S" represents the Single-Stream strategy.

our complete network that combines the HCA module and the SA module in DE-AE produces the best saliency maps.

3) Effectiveness of the Progressive Fusion Strategy: We further examine the efficacy of our Nested Fusion ("Nest") strategy. Specifically, we construct another two network variants, as shown in Fig. 11, where the heterogeneous features are fused in a traditional Dual-Stream ("2-S") or Single-Stream ("1-S") manner. The quantitative results are presented in lines No.8 and No.9 of Table III. Both the "2-S" and "1-S" network variants clearly exhibit significant decreases in both the F_β and E_ξ metrics. This fully demonstrates that our proposed nested fusion strategy helps to progressively fuse two heterogeneous features, avoiding the loss of information caused by the crude combination of traditional methods, and ensuring the integrity and accuracy of the predicted salient objects.

Fig. 12 presents visual comparisons of different fusion strategies. Replacing our "Nest" with a traditional "2-S" or "1-S" fusion strategy often leads to missed objects or falsely highlighted backgrounds. In contrast, our nested fusion yields more accurate predictions for both multiple and SSOs, demonstrating its effectiveness in seamlessly integrating heterogeneous features while avoiding the information loss and mutual interference common in traditional methods.

TABLE IV
ABLATION EXPERIMENTS ON THE SINGLE ENCODER AND DUAL ENCODER. THE BEST RESULTS EQUIPPED WITH VGG OR RESNET ARE, RESPECTIVELY, MARKED IN BOLD FACE

Type	Model Variants	CNN Branch	Flops	Params	ORSSD		EORSSD	
					$S_\alpha\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$M\downarrow$
Single Encoder	SwinT	/	5.21G	377.43M	0.9326	0.0107	0.9245	0.0073
	Conformer	/	26.84G	200.21M	0.9427	0.0088	0.9313	0.0070
	CNN	VGG	30.30G	79.07M	0.9230	0.0134	0.9246	0.0078
		RES	14.78G	139.32M	0.9290	0.0105	0.9253	0.0072
	GCN	/	12.92G	33.96M	0.9149	0.0142	0.9129	0.0101
Dual Encoder	Nest (CNN+GCN)	VGG	154.37G	181.16M	**0.9498**	**0.0077**	**0.9376**	**0.0069**
		RES	30.90G	200.71M	**0.9475**	**0.0078**	**0.9377**	**0.0063**
	2-S (CNN+GCN)	VGG	61.44G	124.16M	0.9398	0.0095	0.9336	0.0071
		RES	18.86G	189.66M	0.9457	0.0085	0.9358	0.0070
	1-S (CNN+GCN)	VGG	57.51G	132.49M	0.9437	0.0090	0.9353	0.0070
		RES	18.81G	192.94M	0.9452	0.0086	0.9368	0.0069

4) Analysis of Different Encoder Configurations: In this part, we provide a comprehensive analysis of the performance and computational complexity of using a single encoder or dual encoders, with the quantitative results reported in Table IV. In the "Single Encoder" setup, we adopted a Swin Transformer (abbreviated as SwinT) [47], a Conformer [63], VGG, ResNet (abbreviated as RES), or our specially designed GCN as a standalone encoder. As shown in Table IV, "Conformer" and "SwinT" achieve the two best performances in the "Single Encoder" configuration due to their strong local and global modeling capabilities. While the single CNN (VGG or ResNet) and GCN achieve slightly inferior performance than the above Transformers do, they exhibit lower computational complexity. Notably, the GCN, a simple network that we constructed without pretraining, is trained directly on either the ORSSD or EORSSD dataset and still achieves commendable performance. In the "Dual Encoder" setup, our final model, "Nest (CNN+GCN)", has two versions: one utilizes VGG as the CNN branch and the other utilizes ResNet as the CNN branch. As shown in Table IV, both two "Nest" generally achieve the two best performances across all datasets. "Nest (VGG+GCN)" has more FLOPs, which is due primarily to the shallow down-sampling strategy of the VGG branch (retaining a 256 $\times$ 256 resolution in the initial layers). In contrast, "Nest (RES+GCN)" reduces the number of FLOPs by 80% (from 154.37G to 30.90G) due to the more aggressive down-sampling strategy of the ResNet branch (reducing the resolution to 64 $\times$ 64 in the early layers). Furthermore, our "Nest (RES+GCN)" and "Conformer" have comparable computational complexity, yet our model achieves superior performance, demonstrating an effective balance between performance and computational complexity. Additionally, we analyze the complexity of using nested fusion in the "Dual Encoder" setup by comparing single-stream (abbreviated as "1-S") and dual-stream (abbreviated as "2-S") approaches. As shown in Table IV, while our Nested Fusion (abbreviated as "Nest (CNN+GCN)") approach

TABLE V
ABLATION EXPERIMENTS FOR THE CONFIGURATIONS (IMAGE PATCH NUMBERS, PATCH NODE NUMBERS AND GCN LAYERS) IN THE E-GCN. THE BEST RESULTS ARE REMARKED IN BOLD FACE

No.	E-GCN			ORSSD		EORSSD	
	patch_num	patch_node	Layers	$S_\alpha\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$M\downarrow$
1	2×2	1	1	0.9473	0.0084	0.9332	0.0070
2	2×2	2	1	0.9482	0.0078	0.9345	0.0071
3	2×2	3	1	**0.9498**	**0.0077**	**0.9376**	**0.0069**
4	2×2	4	1	0.9441	0.0083	0.9365	0.0072
5	4×4	1	1	0.9450	0.0089	0.9343	0.0080
6	4×4	2	1	0.9493	0.0079	0.9369	0.0079
7	4×4	3	1	0.9477	0.0083	0.9333	0.0080
8	4×4	4	1	0.9442	0.0082	0.9331	0.0083
9	2×2	3	1	**0.9498**	0.0077	**0.9376**	**0.0069**
10	2×2	3	2	0.9495	**0.0075**	0.9369	0.0071
11	2×2	3	3	0.9468	0.0085	0.9364	0.0080
12	2×2	3	4	0.9461	0.0088	0.9348	0.0082

has slightly higher computational complexity than "1-S" and "2-S" approaches do, it achieves superior performance. This demonstrates that our progressive nested fusion strategy, with an acceptable increase in complexity, effectively facilitates the fusion of heterogeneous features and efficiently addresses challenges, such as multiscale targets and complex topologies in remote sensing images, validating its superiority.

5) Analysis of Different Configurations In E-GCN: In this ablation part, we analyze the impact of different numbers of image patches, patch nodes, and GCN layers on the performance, as presented in Table V. Rows 1–8 focus on the influence of varying numbers of image patches and patch nodes on performance. Specifically, rows 1–4 report the influence of varying number of patch node on performance for a 2×2 patches setting. The optimal result is attained when the number of patch nodes is set to 3 (row No. 3). Similarly, rows 5–8 report the influence of varying number of patch nodes on performance for a 4×4 patch setting, which shows that a good result is obtained when the number of patch nodes is adjusted to 2 (row No.6). Overall, the optimal performance is achieved for parameter settings of 2×2 patches and 3 patch nodes. Furthermore, we study the effect of varying the number of GCN layers on performance. The results presented in rows 9–12 reveal that the optimal outcome is achieved when only a single GCN layer is employed. However, the MAE metric on the ORSSD dataset is slightly inferior to those of the other settings, and the difference is negligible. Thus, we adopt a 2×2 patch division with 3 patch nodes and 1 GCN layer as the default setting for all the experiments.

V. CONCLUSION

This article delves into the diversity and complementarity of heterogeneous features and introduces a novel CMNFNet, which is composed of an E-VGG encoder subnetwork, an E-GCN encoder subnetwork, and an AECMNFM. The E-VGG encoder is devised to model basic local features. The E-GCN encoder is a graph-based convolution encoder, which is devised to perceive the salient object with different scales or complex topological structures, by conducting parallel message passing over two graph structure data with different receptive fields. Finally, the heterogeneous features extracted from the two encoders are seamlessly fused within the AECMNFM module. This fusion process completes the feature integration in a progressive manner, while simultaneously adapting to eliminate background interference. The extensive experiments and ablation studies robustly showcase the efficacy and superiority of our approach over 16 SOTA models. Although our CMNFNet demonstrates superior performance, the computational complexity of the dual-encoder architecture requires further optimization compared with that of the single-encoder models. In future work, we plan to further optimize the architecture to achieve an effective balance between detection performance and computational efficiency and enhance its real-world applicability.

REFERENCES

[1] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[2] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2941–2959, Oct. 2019.

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

[4] 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.*, 2020, pp. 9410–9419.

[5] J. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8778–8787.

[6] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3080–3089.

[7] M. Xu, P. Fu, B. Liu, H. Yin, and J. Li, "A novel dynamic graph evolution network for salient object detection," *Appl. Intell.*, vol. 52, no. 3, pp. 2854–2871, 2021.

[8] 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.*, vol. 57, no. 11, pp. 9156–9166, Nov. 2019.

[9] Q. Zhang et al., "Dense attention fluid network for salient object detection in optical remote sensing images," *IEEE Trans. Image Process*, vol. 30, pp. 1305–1317, 2021.

[10] 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.*, vol. 60, pp. 1–13, 2022.

[11] L. Zhang, D. Chen, J. Ma, and J. Zhang, "Remote-sensing image superresolution based on visual saliency analysis and unequal reconstruction networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4099–4115, Jun. 2020.

[12] X. Gong, Z. Xie, Y. Liu, X. Shi, and Z. Zheng, "Deep salient feature based anti-noise transfer network for scene classification of remote sensing imagery," *Remote Sens.*, vol. 10, no. 3, pp. 1–24, 2018.

[13] L. Zhang, A. Li, Z. Zhang, and K. Yang, "Global and local saliency analysis for the extraction of residential areas in high-spatial-resolution remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 3750–3763, Jul. 2016.

[14] J. Nie et al., "Scale–relation joint decoupling network for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[15] D. Hong et al., "Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 52–87, Jun. 2021.

[16] G. Li, Z. Liu, D. Zeng, W. Lin, and H. Ling, "Adjacent context coordination network for salient object detection in optical remote sensing images," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 1–13, Jan. 2023.

[17] G. Li, Z. Liu, Z. Bai, W. Lin, and H. Ling, "Lightweight salient object detection in optical remote sensing images via feature correlation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[18] 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.*, 2020, pp. 35–51.

[19] 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.*, vol. 60, pp. 1–15, 2022.

[20] 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.*, vol. 60, pp. 1–13, 2022.

[21] R. Cong, Y. Zhang, L. Fang, J. Li, Y. Zhao, and S. Kwong, "RRNet: Relational reasoning network with parallel multiscale attention for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[22] X. Zhou et al., "Edge-guided recurrent positioning network for salient object detection in optical remote sensing images," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 539–552, Jan. 2023.

[23] Q. Wang, Y. Liu, Z. Xiong, and Y. Yuan, "Hybrid feature aligned network for salient object detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[24] K. Huang, N. Li, J. Huang, and C. Tian, "Exploiting memory-based cross-image contexts for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.

[25] Y. Gu, S. Chen, X. Sun, J. Ji, Y. Zhou, and R. Ji, "Optical remote sensing image salient object detection via bidirectional cross-attention and attention restoration," *Pattern Recognit.*, vol. 164, Aug. 2025, Art. no. 111478.

[26] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: A novel saliency detection framework," *IEEE Trans. Image Process*, vol. 23, no. 5, pp. 1937–1952, May 2014.

[27] S. Wang, M. Wang, S. Yang, and K. Zhang, "Salient region detection via discriminative dictionary learning and joint Bayesian inference," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 5, pp. 1116–1129, 2018.

[28] Y. Yuan, C. Li, J. Kim, W. Cai, and D. D. Feng, "Reversion correction and regularized random walk ranking for saliency detection," *IEEE Trans. Image Process*, vol. 27, no. 3, pp. 1311–1322, May 2018.

[29] J.-G. Yu, J. Zhao, J. Tian, and Y. Tan, "Maximal entropy random walk for region-based visual saliency," *IEEE Trans. Cybern.*, vol. 44, no. 9, pp. 1661–1672, Sep. 2014.

[30] C. Li, Y. Yuan, W. Cai, Y. Xia, and D. D. Feng, "Robust saliency detection via regularized random walks ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2710–2717.

[31] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, "Salient object detection via structured matrix decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 818–832, Apr. 2017.

[32] R. Cong, J. Lei, H. Fu, F. Porikli, Q. Huang, and C. Hou, "Video saliency detection via sparsity-based reconstruction and propagation," *IEEE Trans. Image Process*, vol. 28, no. 10, pp. 4819–4831, Oct. 2019.

[33] M. Xu, B. Liu, P. Fu, J. Li, and Y. H. Hu, "Video saliency detection via graph clustering with motion energy and spatiotemporal objectness," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2790–2805, Nov. 2019.

[34] M. Xu, B. Liu, P. Fu, J. Li, Y. H. Hu, and S. Feng, "Video salient object detection via robust seeds extraction and multi-graphs manifold propagation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 2191–2206, Jul. 2020.

[35] S. Mohammadi, M. Noori, A. Bahri, S. G. Majelan, and M. Havaei, "CAGNet: Content-aware guidance for salient object detection," *Pattern Recognit.*, vol. 103, pp. 1–12, Jul. 2020.

[36] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5961–5970.

[37] M. Huang, Z. Liu, L. Ye, X. Zhou, and Y. Wang, "Saliency detection via multi-level integration and multi-scale fusion neural networks," *Neurocomputing*, vol. 364, pp. 310–321, Oct. 2019.

[38] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 714–722.

[39] S. Chen, B. Wang, X. Tan, and X. Hu, "Embedding attention and residual network for accurate salient object detection," *IEEE Trans. Cybern.*, vol. 50, no. 5, pp. 2050–2062, May 2020.

[40] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 1913–1927, Aug. 2020.

[41] P. Zhang, W. Liu, H. Lu, and C. Shen, "Salient object detection with lossless feature reflection and weighted structural loss," *IEEE Trans. Image Process*, vol. 28, no. 6, pp. 3048–3060, Jun. 2019.

[42] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1623–1632.

[43] Z. Tu, Y. Ma, C. Li, J. Tang, and B. Luo, "Edge-guided non-local fully convolutional network for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 582–593, Feb. 2021.

[44] J. Jin, Q. Jiang, Q. Wu, B. Xu, and R. Cong, "Underwater salient object detection via dual-stage self-paced learning and depth emphasis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 3, pp. 2147–2160, Mar. 2025.

[45] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[46] Y. Mao et al. "Transformer transforms salient object detection and camouflaged object detection." Apr. 2021. [Online]. Available: https://arxiv.org/abs/2402.18922

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

[48] Y. Ge, Q. Zhang, T.-Z. Xiang, C. Zhang, and H. Bi, "TCNet: Co-salient object detection via parallel interaction of transformers and CNNs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2600–2615, Jun. 2023.

[49] C. Xie, C. Xia, M. Ma, Z. Zhao, X. Chen, and J. Li, "Pyramid grafting network for one-stage high resolution saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11707–11716.

[50] J. Cheng, Z. Wu, S. Wang, C. Demonceaux, and Q. Jiang, "Bidirectional collaborative Mentoring network for marine organism detection and beyond," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6595–6608, Nov. 2023.

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

[52] Q. Jiang, J. Cheng, Z. Wu, R. Cong, and R. Timofte, "High-precision dichotomous image segmentation with frequency and scale awareness," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 5, pp. 8619–8631, May 2025.

[53] M. Xu, P. Fu, B. Liu, and J. Li, "Multi-stream attention-aware graph convolution network for video salient object detection," *IEEE Trans. Image Process*, vol. 30, pp. 4183–4197, 2021.

[54] Q. Zhang, S. Wang, X. Wang, Z. Sun, S. Kwong, and J. Jiang, "Geometry auxiliary salient object detection for light fields via graph neural networks," *IEEE Trans. Image Process*, vol. 30, pp. 7578–7592, 2021.

[55] B.-W. Yin and Z. Lin, "Exploring salient object detection with adder neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, Apr. 2025, pp. 9490–9498.

[56] J. Zhao, Y. Jia, L. Ma, and L. Yu, "Adaptive dual-stream sparse transformer network for salient object detection in optical remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 5173–5192, 2024.

[57] B. Liang and H. Luo, "MEANet: An effective and lightweight solution for salient object detection in optical remote sensing images," *Exp. Syst Appl*, vol. 238, pp. 1–14, Mar. 2024.

[58] Y. Li and A. Gupta, "Beyond grids: Learning graph representations for visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[59] D.-P. 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.*, 2017, pp. 4558–4567.

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

[61] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 698–704.

[62] 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.*, 2012, pp. 733–740.

[63] S. Kanwal and I. A. Taj, "Incomplete RGB-D salient object detection: Conceal, correlate and fuse," *Pattern Recognit.*, vol. 155, pp. 1–15, Nov. 2024.