GLC-Net: Global-Local Collaborative Network for Remote Sensing Image Segmentation

WEI Kan[1,2], *LI Ling*[1], *LIANG Shilin*[1], *WEN Zongguo*[1*]

1. School of Environment, Tsinghua University, Beijing 100084, P. R. China;
2. Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, P. R. China

(Received 11 August 2025; revised 30 September 2025; accepted 13 October 2025)

Abstract: Intelligent interpretation of high-resolution remote sensing imagery is a fundamental challenge in aerospace information processing. Complex ground environments such as construction and demolition (C&D) waste landfills exemplify the need for robust segmentation models that can handle diverse spatial and spectral patterns. Conventional convolutional neural networks (CNNs) are limited by their local receptive fields, whereas Transformer-based architectures often lose fine spatial detail, resulting in incomplete delineation of heterogeneous remote sensing targets. To address these issues, we propose a global-local collaborative network (GLC-Net), which is designed for intelligent remote sensing image segmentation. The model integrates an efficient Transformer block to capture global dependencies and a local enhancement block to refine structural details. Furthermore, a multi-scale spatial aggregation and enhancement (MSAE) module is introduced to strengthen contextual representation and suppress background noise. Deep supervision facilitates hierarchical feature learning. Experiments on two high-resolution remote sensing datasets (Changping and Daxing) demonstrate that GLC-Net surpasses state-of-the-art baselines by 1.5%—3.2% in mean intersection over union (mIoU), while achieving superior boundary precision and semantic consistency. These results confirm that global-local collaborative modeling provides an effective pathway for intelligent remote sensing image segmentation in aerospace environmental monitoring.

Key words: remote sensing imagery; deep learning; vision transformer; landfill; segmentation

CLC number: TP751.1 **Document code:** A **Article ID:** 1005-1120(2025)05-0565-12

0 Introduction

The accelerating pace of global urbanization has triggered a substantial increase in construction and demolition (C&D) waste. This type of waste not only occupies valuable land resources but also poses persistent environmental risks through the potential leaching of hazardous substances, including heavy metals and asbestos, into soil and groundwater systems. Consequently, effective monitoring and assessment of C&D waste landfill sites are critical for advancing sustainable urban development and ensuring environmental protection[1].

Traditional monitoring methods, such as manual field surveys, are often hampered by high operational costs, low efficiency, and limited spatial coverage. In this context, remote sensing technology emerges as a powerful alternative, providing wide-area, repeatable, and non-intrusive observation capabilities. The advent of high-resolution satellite imagery has further enabled a more consistent and detailed capture of the spatial distribution and temporal dynamics of landfill sites. When integrated with deep learning techniques, remote sensing facilitates the automated and accurate extraction of landfill-related information, thereby significantly enhancing monitoring efficiency and supporting informed decision-making in waste management[2].

In recent years, deep learning has become the

*Corresponding author, E-mail address: wenzg@tsinghua.edu.cn.

How to cite this article: WEI Kan, LI Ling, LIANG Shilin, et al. GLC-Net: Global-local collaborative network for remote sensing image segmentation[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2025,42(5):565-576.

http://dx.doi.org/10.16356/j.1005-1120.2025.05.001

dominant paradigm for interpreting remote sensing imagery. Semantic segmentation models based on convolutional neural networks (CNNs) have achieved notable success in landfill detection tasks. For instance, UNet[3] employs a symmetric encoder-decoder architecture with skip connections, effectively recovering spatial details that are lost during the downsampling process. DeepLabV3+[4] extends this framework by incorporating atrous convolutions and an atrous spatial pyramid pooling (ASPP) module, which capture multi-scale contextual information and enhance the recognition of irregular targets. While these CNN-based methods exhibit strong capabilities in local feature extraction, their capacity to represent complex large-scale landfill environments remains constrained by the inherent locality of convolutional operations.

The inherent limitation of CNNs lies in the intrinsic locality of their convolutional kernels. Although highly effective at capturing local neighborhood patterns, CNNs typically struggle to model long-range, global dependencies across an entire scene. This limitation consequently restricts a model's ability to comprehend large-scale spatial layouts and complex contextual relationships, such as the correlations among dispersed waste piles, adjacent vegetation, and engineered structures within a landfill site. In response to this shortcoming, Transformer-based architectures have been increasingly adopted for semantic segmentation. For example, SegFormer[5] replaces standard convolutions with a hierarchical Transformer encoder and a lightweight multilayer perceptron (MLP) decoder. By leveraging the self-attention mechanism, it effectively captures global contextual dependencies while maintaining computational efficiency. Similarly, UNetFormer[6] integrates the powerful global modeling capabilities of Transformers into a U-shaped network, hybridizing self-attention with convolutional operations to strike a balance between local detail preservation and global context integration. While these Transformer-based models have demonstrated compelling performance in natural image segmentation and showed considerable promise for remote sensing applications, their specific potential and adaptation for the task of landfill detection remain largely underexplored. Moreover, the fusion of global and local features is crucial for tasks like semantic segmentation. STransFuse[7] combines a swin Transformer with a CNN to capture both global context and local spatial features, but its reliance on large datasets limits its remote sensing applicability. ST-UNet[8] integrates a Transformer with a UNet, using spatial interaction and feature compression modules to enhance segmentation, particularly for small-scale objects. However, existing hybrid architectures face limitations in complex remote sensing environments like landfill detection. In contrast, the global-local collaborative network (GLC-Net) integrates global context and local details through a Global Block for long-range dependencies and a Local Block for fine-grained details. Unlike ST-UNet and DeSwin-S[9], GLC-Net does not require parallel CNN-Transformer branches. The local block serially refines global information in a simple and efficient manner, resulting in a more lightweight model. Moreover, GLC-Net extracts global features directly during the feature extraction stage, rather than relying on self-attention only in the decoder as in UNetFormer, enabling more comprehensive feature representation and improved robustness for challenging tasks.

Beyond the global context modeling challenge, another critical issue is the effective representation of the highly heterogeneous spatial composition characteristic of landfill sites. In high-resolution imagery, such areas typically exhibit irregular geometries, complex textural patterns, and multiple coexisting land-cover types, including exposed waste heaps, operational facilities, access roads, and patches of vegetation. Prevailing segmentation models often resort to simplistic strategies, such as direct concatenation or element-wise addition, when fusing multi-scale features. These approaches are frequently inadequate for capturing the intricate spatial arrangements and nuanced semantic relationships among the diverse components within a landfill. Consequently, such representational shortcomings commonly lead to misclassifications in cluttered regions and imprecise boundary delineation in the re-

sulting segmentation maps.

To address the aforementioned challenges, this study proposes a novel deep learning framework, termed GLC-Net, for the accurate and robust semantic segmentation of CD waste landfills. The principal contributions of this work are summarized as follows:

(1) We propose a global-local collaborative framework that synergistically leverages CNNs for fine-grained detail extraction and Transformers for global context modeling. This design mitigates the limitations of local receptive fields in conventional CNN-based models and enables a more holistic representation of complex landfill scenes.

(2) To effectively model the heterogeneity of landfill areas, we design a multi-scale spatial aggregation and enhancement (MSAE) module. This module is engineered to explicitly integrate multi-scale features and strengthen spatial relation learning, thereby achieving improved boundary delineation and key region recognition.

(3) We conduct extensive experiments on the Changping and Daxing datasets. The proposed method demonstrates superior performance, outperforming state-of-the-art baselines across multiple metrics and confirming its robustness and practical potential.

1 Methods

To accurately identify and segment construction waste landfills in high-resolution remote sensing imagery, we propose the novel GLC-Net as shown in Fig.1. The core of GLC-Net lies in its integration of the local detail extraction capability of CNNs with the global context modeling power of Transformers. This synergy is further augmented by the specially designed MSAE module to enhance feature representation for complex landfill scenes. This section delineates the overall architecture of GLC-Net, elaborates on the design principles of its core modules, and explains the loss function strategy employed for robust semantic segmentation.

1.1 Global-local collaborative modeling framework

The proposed GLC-Net follows an encoder-decoder framework, architected to jointly capture global context and local details for precise landfill segmentation, as depicted in Fig.1, where H, W, C, D and N denote the image height, width, channel dimension, the dimension of the feature map, and the number of classes, respectively. The network mainly consists of the efficient Transformer (ET)

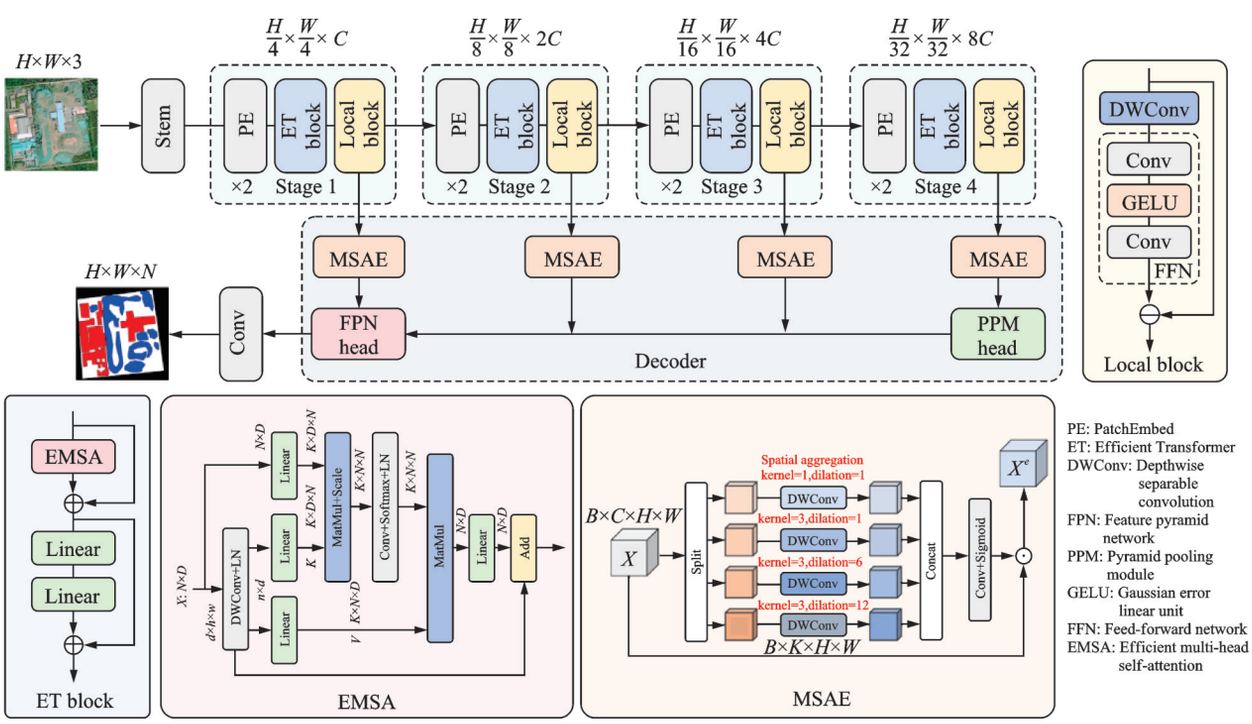

Fig.1 Pipeline of the proposed GLC-Net

block and the MSAE module. Given an input image $\boldsymbol{I} \in \mathbf{R}^{H \times W \times 3}$, it is first processed by a stem layer composed of convolutional operations that perform initial downsampling and channel expansion, yielding a base feature map, shown as

$$\boldsymbol{F}_0 = \mathrm{Conv}_{\mathrm{stem}}(\boldsymbol{I}) \in \mathbf{R}^{H/4 \times W/4 \times C} \tag{1}$$

This feature map serves as the input to a hierarchical encoder comprising four stages. At each stage i, the spatial resolution of the feature map is reduced by a factor of two while the channel dimension is doubled, yielding $\boldsymbol{F}_i \in \mathbf{R}^{\frac{H}{2^i+1} \times \frac{H}{2^i+1} \times 2^{i-1}C}$.

Within each stage, the network sequentially employs an ET block[10] and a local block. The ET block leverages an EMSA mechanism to model long-range dependencies, thereby capturing the global contextual information essential for identifying spatially dispersed landfill regions. Formally, the attention operation is expressed as

$$\boldsymbol{F}_i^{\mathrm{ET}} = \mathrm{EMSA}(\boldsymbol{F}_i^{\mathrm{in}}) \tag{2}$$

$$\boldsymbol{H}_k = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_k \boldsymbol{K}_k^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{3}$$

$$\mathrm{EMSA}(\boldsymbol{F}) = \mathrm{Concat}(\boldsymbol{H}_1, \boldsymbol{H}_2, \cdots, \boldsymbol{H}_h)\boldsymbol{W}^O \tag{4}$$

where $\boldsymbol{F}_i^{\mathrm{in}}$ is the input to stage i; $\boldsymbol{Q}_k$, $\boldsymbol{K}_k$, and $\boldsymbol{V}_k$ are the queries, keys, and values for the kth attention head; h is the number of heads, d_k the head dimension, and $\boldsymbol{W}^O$ a projection matrix[11]. The ET block ensures that distant regions within the landfill can interact, establishing a globally coherent representation.

Following the ET block, the local block refines the globally contextualized features using depthwise separable convolutions and residual connections, shown as

$$\boldsymbol{F}_i = \boldsymbol{F}_i^{\mathrm{ET}} + \mathrm{Conv}_{\mathrm{dw}}(\boldsymbol{F}_i^{\mathrm{ET}}) \tag{5}$$

Eq.(5) enhances local textures and neighborhood relations while preserving global semantic context. Unlike conventional convolutional refinement modules that process features in isolation, the local block sequentially integrates global and local information in a lightweight manner, allowing more effective contextual adaptation with minimal computational overhead. Its structure ensures targeted enhancement of fine-grained details without disrupting the globally aggregated features. By applying this global-local sequence at every stage, the encoder produces a multi-scale feature pyramid $\{\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3, \boldsymbol{F}_4\}$, where features at each level progressively encode richer semantics with coarser spatial resolution. These hierarchical features integrate global dependencies and local details, forming a robust foundation for the decoder to generate high-precision segmentation outputs.

1.2 MSAE module

To improve the representational capacity of hierarchical features extracted by the hybrid CNN-Transformer backbone, we propose the MSAE module. Integrated at the output of each backbone stage, the MSAE module explicitly captures multi-scale spatial dependencies and adaptively recalibrates channel-wise feature responses. By emphasizing semantically salient regions while suppressing background noise, the module enhances the discriminability of features for complex landfill scenes, facilitating more accurate delineation of irregular structures and heterogeneous land cover types.

Let the feature map output from the local block of stage i in the backbone be denoted as $\boldsymbol{F}_i \in \mathbf{R}^{H_i \times W_i \times C_i}$. This feature map serves as the input to the corresponding MSAE module. First, the feature map is evenly split along the channel dimension into four sub-feature maps $\{G_j\}_{j=1}^4$, and each of size is $H_i \times W_i \times C_i/4$. This channel splitting strategy allows parallel processing of distinct sub-feature maps, facilitating the capture of heterogeneous spatial patterns.

Each sub-feature map G_j is then processed through an independent branch composed of a depthwise separable convolution, denoted as $\mathcal{D}(\cdot)$. By employing different kernel sizes or dilation rates across the branches, the module extracts multi-scale spatial features from the same input map. The transformed feature map from each branch is given by

$$\boldsymbol{H}_j = \mathcal{D}(G_j) \tag{6}$$

The multi-scale features from all four branches are concatenated along the channel dimension to form an aggregated feature map $\boldsymbol{F}_{\mathrm{agg}}^i \in \mathbf{R}^{H_i \times W_i \times C_i}$,

shown as

$$F_{\text{agg}}^{i} = \text{Concat}\left[H_1, H_2, H_3, H_4\right] \tag{7}$$

This aggregated representation, containing rich multi-scale spatial context, is further processed to generate a 2D spatial attention map $M_s \in \mathbf{R}^{H_i \times W_i \times 1}$ through a 1×1 convolution followed by a Sigmoid activation $\sigma(\cdot)$, shown as

$$M_s = \sigma(\text{Conv}_{1\times1}(F_{\text{agg}})) \tag{8}$$

Finally, the original input feature map F_i is refined by the spatial attention map through element-wise multiplication $\otimes$ and a residual connection, producing the output F_i' of the MSAE module, shown as

$$F_i' = F_i + (F_i \otimes M_s) \tag{9}$$

By incorporating MSAE modules at every backbone stage, the network progressively refines the hierarchical features, enhancing their discriminative power for the decoder. This design allows the model to accurately capture both fine-grained details and global spatial dependencies in complex landfill scenes, resulting in improved segmentation performance.

1.3 Decoder and deep supervision

The decoder is constructed as a multi-scale fusion architecture to integrate hierarchical features from all stages of the backbone. The deepest feature map F_4' is first processed through a PPM[12] to capture contextual information at multiple scales, enhancing the global receptive field. The three stage-wise feature maps, $\{F_i'\}_{i=2}^{4}$ are then fused through a FPN[13], which leverages top-down pathways and lateral connections to combine high-level semantic cues with detailed spatial information. The final segmentation map is obtained by applying a convolution to the FPN output, shown as

$$\hat{Y}_{\text{final}} = \text{Conv}_{1\times1}(\text{FPN}(F_1', F_2', F_3', \text{PPM}(F_4'))) \tag{10}$$

To facilitate deep supervision, the four backbone features before FPN are individually passed through convolutions to generate auxiliary multi-scale segmentation maps $\{\hat{Y}_i\}_{i=1}^{4}$. These auxiliary predictions provide intermediate gradient signals during training, guiding the backbone to learn more discriminative multi-scale features and improving convergence, boundary delineation, and robustness against heterogeneous spatial structures. All outputs are rescaled via bilinear interpolation to match the input resolution $H \times W$.

The total training loss ($\mathcal{L}$) combines Dice and Focal losses ($\mathcal{L}_{\text{dice}}$ and $\mathcal{L}_{\text{focal}}$) and incorporates contributions from both the main and auxiliary outputs ($\mathcal{L}_{\text{main}}$ and $\mathcal{L}_{\text{aux}}$) in a weighted manner, shown as

$$\mathcal{L} = \mathcal{L}_{\text{main}} + \alpha \times \sum_{i=1}^{4} \mathcal{L}_{\text{aux}}^{(i)} \tag{11}$$

$$\mathcal{L}_{\text{main}} = \mathcal{L}_{\text{dice}}\left(\hat{Y}_{\text{final}}, Y\right) + \mathcal{L}_{\text{focal}}\left(\hat{Y}_{\text{final}}, Y\right) \tag{12}$$

$$\mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{dice}}\left(\hat{Y}_i, Y\right) + \mathcal{L}_{\text{focal}}\left(\hat{Y}_i, Y\right) \tag{13}$$

where Y denotes the ground-truth segmentation map and α is set to 0.1 during training. By combining multiscale supervision with hierarchical feature fusion, the decoder effectively balances local detail preservation and global contextual understanding, producing highly accurate segmentation results for complex landfill scenes.

2 Experiment and Analysis

2.1 Experimental datasets

The proposed method was evaluated on two high-resolution datasets[14]: Changping (CP) and Daxing (DX). The Changping dataset comprises 1 368 images of size 512 pixel × 512 pixel, acquired from the GF-2 satellite, which provides 0.8 m panchromatic and 3.2 m multispectral imagery with approximately 80 cm ground sample distance (GSD). The Daxing dataset contains 2 285 images of the same size, obtained from Google Earth via the Google API, with a spatial resolution of approximately 50 cm GSD.

Both datasets were split into training and testing sets at an 8∶2 ratio, yielding 1 094 training and 274 testing images for Changping, and 1 828 training and 457 testing images for Daxing. Each image is annotated with four semantic categories: Background, vacant landfillable area, engineering facility area, and waste dumping area. These classes reflect the key components of construction and demolition landfill sites and provide a fine-grained basis for evaluating semantic segmentation performance.

The combination of high spatial resolution, di-

verse land-cover types, and clearly defined categories makes these datasets suitable benchmarks for assessing the accuracy and robustness of landfill detection methods.

2.2 Experimental implementation

We evaluated our model using three established semantic segmentation metrics: Overall accuracy (OA), mean intersection over union (mIoU), and mean F_1 score (mF1). OA measures global pixel-wise classification accuracy, while mIoU assesses per-class region overlap between predictions and ground truth. mF1 balances precision and recall across all categories, providing a comprehensive performance assessment through complementary perspectives on segmentation quality and class-wise consistency.

All experiments were implemented in PyTorch on a single NVIDIA RTX 4090 GPU. The model was trained for 100 epochs with a batch size of 32, using the AdamW optimizer with a learning rate of 0.001 and weight decay of 0.000 5, enhanced by a warm-up poly learning rate scheduler.

2.3 Performance comparison

To comprehensively evaluate the effectiveness of the proposed GLC-Net, we conduct comparisons with a set of representative baseline methods spanning three major technical paradigms. These include conventional CNNs such as DeepLab-V3+ and U-Net lightweight CNN-based architectures including A2-FPN[15], ABCNet-E[16] and BA-Net[17], as well as Transformer-based models such as SegFormer[5], UNetFormer, DeSwin-S, and CMTFNet[18].

The quantitative results, summarized in Table 1 and Table 2, demonstrate that GLC-Net consistently outperforms all baseline methods across both datasets, where BA represents the background area, VLA the vacant landfillable area, EFA the engineering facility area, and WDA the waste dumping area. On the CP dataset (Table 1), our method achieves outstanding performance with 92.28% of OA, 80.08% of mIoU, and 88.75% of mF1, surpassing the second-best approach by significant margins of 2.22%, 6.59%, and 3.74%, respectively. These substantial improvements are particularly notable when compared against traditional CNN architectures and their lightweight variants. The superior performance validates the effectiveness of our proposed global-local hybrid modeling paradigm, which synergistically combines the fine-grained detail capture capability of CNNs with the long-range contextual reasoning strength of Transformers. This complementary integration enables the network to extract more discriminative features across diverse spatial scales and complex landfill scenarios, addressing the challenging nature of the CP dataset.

Table 1 Performance comparison of the CP dataset %

Method	BG	VLA	EFA	WDA	OA	mIoU	mF1
A2-FPN	90.23	71.21	80.87	55.86	83.70	61.03	74.54
ABCNet-E	93.14	73.67	85.61	70.50	87.14	68.69	80.73
BA-Net	**95.23**	84.89	89.51	79.15	91.69	77.79	87.20
DeSwin-S	92.78	80.54	86.72	78.44	89.02	73.76	84.62
DeepLab-V3+	91.62	77.92	84.41	76.83	87.32	70.94	82.69
SegFormer	93.88	82.96	87.23	75.97	90.06	74.49	85.01
U-Net	92.96	80.86	85.26	71.91	88.54	71.29	82.75
UNetFormer	92.63	79.09	85.21	74.51	88.32	71.32	82.86
CMFTNet	94.67	84.97	88.38	80.94	91.28	77.72	87.94
GLC-Net	95.00	**85.66**	**91.03**	**83.32**	**92.28**	**80.08**	**88.75**

Note: The bold and underline values indicate the best and second-best results in each column, respectively.

Table 2 Performance comparison of the DX dataset %

Method	BG	VLA	EFA	WDA	OA	mIoU	mF1
A2-FPN	95.6	89.89	82.32	80.15	91.42	77.51	86.99
ABCNet-E	94.66	85.78	74.03	74.05	88.35	70.64	82.13
BA-Net	95.96	89.92	80.89	83.12	91.56	78.24	87.47
DeSwin-S	95.97	93.40	88.70	87.46	93.90	84.32	91.38
DeepLab-V3+	97.13	95.05	90.78	91.26	95.44	88.01	93.55
SegFormer	97.45	95.04	91.07	91.16	95.59	88.23	93.68
U-Net	97.26	94.97	90.93	91.92	95.49	88.37	93.77
UNetFormer	96.66	93.87	89.01	89.44	94.50	85.77	92.25
CMFTNet	**97.52**	94.84	90.62	91.09	95.48	87.96	93.52
GLC-Net	97.48	**95.35**	**91.21**	**92.67**	**95.80**	**89.10**	**94.18**

Note: The bold and underline values indicate the best and second-best results in each column, respectively

For the DX dataset (Table 2), GLC-Net maintains its performance superiority, attaining 95.80% of OA, 89.10% of mIoU, and 94.18% of mF1 while consistently leading across all evaluation metrics. Although Transformer-based methods such as SegFormer and UNetFormer demonstrate competitive performance on this high-resolution dataset,

our approach exhibits distinct advantages in handling complex scenarios featuring highly irregular waste pile distributions and mixed land-cover types. The robust performance underscores the critical importance of simultaneously maintaining global context awareness and local detail sensitivity for accurate feature extraction in challenging remote sensing environments.

Further analysis indicates that the varying performance of BA-Net across the two datasets mainly reflects differences in model robustness rather than dataset characteristics alone. BA-Net's lightweight CNN design depends heavily on local texture cues, which work well in the CP dataset due to its regular landfill structures and homogeneous backgrounds. However, this reliance becomes a limitation when facing the DX dataset's higher spatial heterogeneity and spectral variations caused by different sensors and resolutions. Compared with CMTFNet, which primarily focuses on multi-scale feature fusion through Transformer encoder, GLC-Net emphasizes a more efficient global-local interaction to enhance spatial consistency and contextual understanding. In contrast, GLC-Net maintains consistent accuracy across both datasets because its global-local collaborative mechanism adaptively balances fine-grained spatial detail with global contextual understanding. By integrating long-range dependency modeling with local feature enhancement, GLC-Net demonstrates stronger generalization and robustness under diverse imaging and scene conditions, an essential capability for real-world remote sensing applications.

Qualitative visualizations shown in Fig.2 further corroborate these findings. GLC-Net generates segmentation maps with sharper boundaries, reduced fragmentation, and better internal consistency within each semantic class. This improvement is attributable not only to the global-local hybrid backbone but also to the MSAE module, which effectively strengthens the representation of complex spatial patterns. In contrast, other methods often produce blurred edges, fragmented regions, or misclassifications in areas with intricate geometry, underscoring the benefits of our explicit spatial information enhancement.

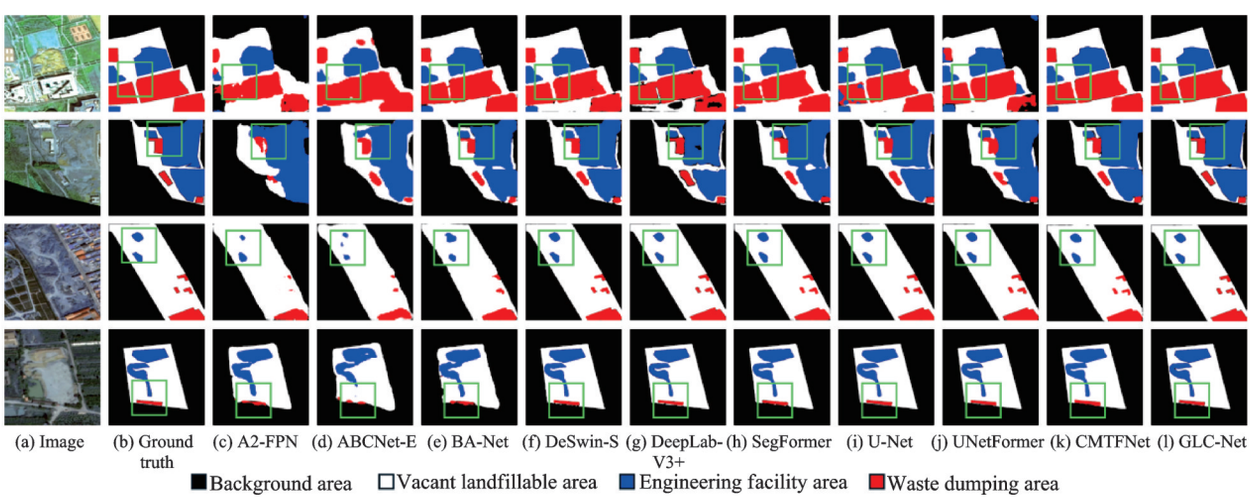

Fig.2 Visual comparison on CP (rows 1—2) and DX (rows 3—4) datasets

Collectively, these results demonstrate that GLC-Net can simultaneously leverage global contextual information and finely detailed local features, delivering more accurate and robust segmentation for construction waste landfill detection.

2.4 Ablation study

To systematically investigate the contributions of the key components in GLC-Net, we conducted ablation experiments on the CP and DX datasets, focusing on three modules: The local block (LB) in each encoder stage, the MSAE module, and the deep supervision (DS) in the decoder. Table 3 summarizes the quantitative results in terms of OA, mIoU, and mF1.

When all three components are included, the

Table 3 Ablation study of the proposed model's components on CP and DX datasets %

Component			CP			DX		
LB	MSAE	DS	OA	mIoU	mF1	OA	mIoU	mF1
√	√	√	**92.28**	**80.08**	**88.75**	**95.80**	**89.10**	**94.18**
	√	√	91.88	78.77	87.96	95.25	87.68	93.26
√		√	91.41	77.53	87.05	95.19	87.41	93.20
√	√		92.03	79.24	88.19	95.48	88.31	93.73
√			91.17	76.95	86.69	94.28	84.94	91.73
	√		91.51	77.75	87.21	95.10	87.17	93.06
		√	90.93	75.93	85.94	93.76	83.42	90.79

Note: The checkmark √ indicates that the corresponding module is included in the configuration.

network achieves its best performance, with 92.28% of OA, 80.08% of mIoU, and 88.75% of mF1 on CP detaset, and 95.80% of OA, 89.10% of mIoU, and 94.18% of mF1 on DX dataset. Removing LB while retaining MSAE and DS results in a drop of 0.40% of OA and 1.11% of mIoU on CP dataset, indicating that local feature enhancement in the encoder is important for capturing fine-grained spatial details. Similarly, excluding MSAE while keeping LB and DS reduces performance by 0.87% of mIoU on CP dataset and 1.69% of mIoU on DX dataset, demonstrating the effectiveness of the proposed module in modeling multi-scale spatial relationships and refining feature representations.

The impact of deep supervision is also evident. Removing DS while keeping LB and MSAE leads to slight declines in OA and mF1 (e.g., 0.05% of OA and 0.56% of mF1 drop on CP dataset), confirming that multi-scale auxiliary supervision facilitates stable training and encourages the network to learn more discriminative intermediate representations. Configurations with only a single module (LB, MSAE, or DS) exhibit the lowest performance, highlighting that each component contributes complementarily to the overall accuracy and segmentation quality.

The ablation results demonstrate that LB, MSAE, and DS contribute complementarily: LB enhances local detail extraction, MSAE captures complex spatial relationships, and DS guides multi-scale feature learning. Together, they enable precise boundary delineation, robust segmentation of heterogeneous regions, and improved overall accuracy in landfill scenes.

To assess the rationality of the auxiliary branch weight in the hybrid loss, we conducted an ablation experiment by varying the auxiliary loss coefficient α from 0.1 to 1.0. The results, presented in Table 4, show that the model achieves the best performance when $\alpha = 0.1$ on both CP and DX datasets. A small auxiliary weight allows the auxiliary branch to provide complementary supervision without dominating the optimization process of the main segmentation branch. As α increases, the auxiliary loss exerts excessive influence, causing the network to overfit local cues and deviate from the optimal global segmentation objective. When $\alpha \geqslant 0.5$, the model's performance begins to decline, and a further increase to $\alpha = 1.0$ leads to noticeable degradation. These findings confirm that a lower auxiliary supervision strikes a better balance between stability and effectiveness in joint optimization.

Table 4 Ablation study on the auxiliary branch weight

α	Dataset	OA/%	mIoU/%	mF1/%
0.1	CP	**92.28**	**80.08**	**88.75**
	DX	**95.80**	**89.10**	**94.18**
0.25	CP	91.94	79.32	88.02
	DX	95.60	88.82	93.78
0.5	CP	91.51	78.63	87.65
	DX	95.25	87.98	93.20
0.75	CP	91.42	78.26	87.34
	DX	95.11	87.49	93.02
1.0	CP	90.78	77.12	86.41
	DX	94.55	86.63	91.95

2.5 Heatmap visualization

To gain deeper insights into the internal workings of GLC-Net and to qualitatively validate the effectiveness of its key modules, we employed gradient-weighted class activation mapping (Grad-CAM)[19] to visualize feature representations. Grad-CAM generates heatmaps that highlight the most important regions in an input image contributing to a specific class prediction. By examining these heatmaps at different depths within the network, we can observe the evolution of feature refinement across stages.

As illustrated in Fig.3, we selected four representative samples and generated activation maps for three primary classes, focusing on four key stages of feature processing: A—before entering the local block, B—before the MSAE module, C—before the PPM head, and D—after the PPM head. Each pixel value in the heat map represents the relative contribution score to the final class logit; higher values denote more informative regions for the category of interest.

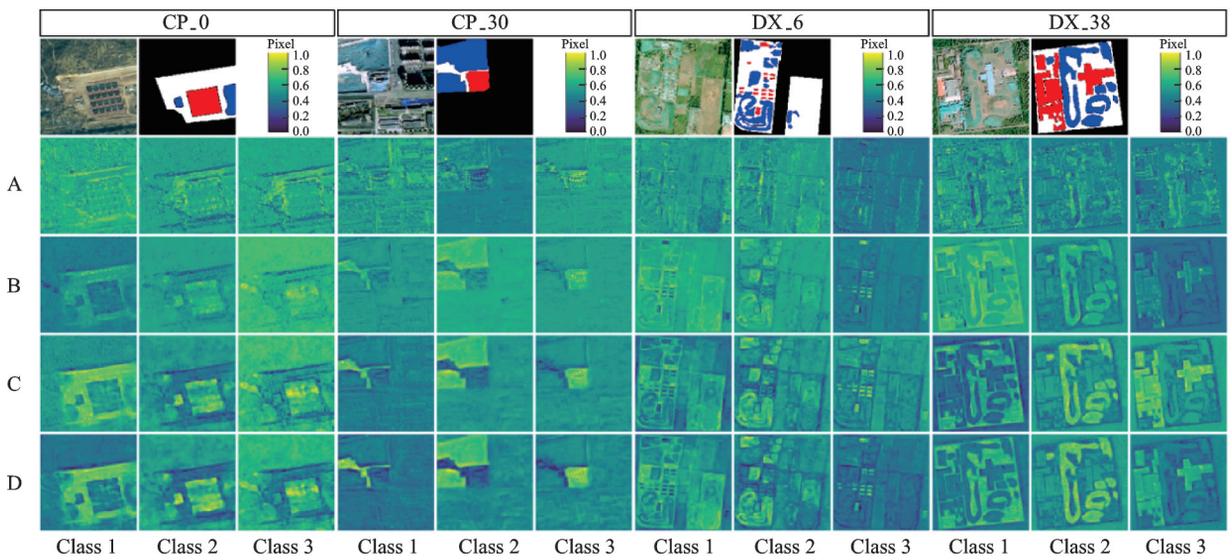

Fig.3 Heatmap visualization of each stage

At stage A, the initial features provide a rough localization of target regions. The highlighted areas are diffuse and contain considerable background noise, indicating that early features encode coarse positional information but exhibit limited semantic discriminability.

Stage B shows the effect of the local block. Activation regions become more concentrated, and minor noise is suppressed. For example, in DX_38, the outlines of structures are clearer compared to stage A, demonstrating that the local block enhances local textures and fine-grained details. However, activations remain fragmented, and global target integrity is not yet fully captured.

After passing through the MSAE module (stage C), previously scattered and disjointed activations are aggregated into semantically coherent and spatially continuous regions. In CP_30, scattered activations at stage B merge at stage C to form contiguous shapes closely matching the ground truth. This highlights the MSAE module's role in integrating multi-scale spatial context, substantially improving semantic consistency and holistic perception.

Finally, at stage D, the PPM head refines the features further, producing highly focused and precise activations. The module enhances global context understanding, accurately delineating object boundaries and suppressing background responses.

Overall, the Grad-CAM visualizations reveal a clear pattern: Features evolve from coarse and diffuse in the early stage, to more localized and structured after the local block, to spatially coherent and semantically enriched after MSAE, and finally to highly precise and globally consistent representations after the PPM head. This progressive refinement underscores the pivotal role of the MSAE module in aggregating multi-scale context, preserving target integrity, and enhancing discriminative power, which collectively supports the superior segmentation performance of GLC-Net in complex landfill scenes.

2.6 Efficiency analysis

To evaluate the practical utility of GLC-Net, we compare its parameters computational cost, and segmentation performance against several state-of-

Table 5 Comparison of parameters and FLOPs of GLC-Net with several SOTA models

Method	Parameter/10^6	FLOP/10^9
A2-FPN	22.82	83.65
ABCNet-E	13.36	30.98
BA-Net	12.69	25.84
DeSwin-S	45.61	93.82
DeepLab-V3+	40.34	93.61
SegFormer	13.73	23.98
U-Net	19.88	123.44
UNetFormer	11.68	23.48
CMTFNet	30.07	66.12
GLC-Net	16.19	24.16

the-art (SOTA) models as shown in Table 5. From Table 5 we can see that, GLC-Net achieves 16.19×10^6 parameters and 24.16×10^8 FLOPs, establishing a robust equilibrium between computational efficiency and segmentation accuracy. Despite having fewer parameters and lower computational cost compared to models such as U-Net and DeepLab-V3+, GLC-Net consistently outperforms these models in mIoU and mF1, highlighting its superior capability in addressing complex remote sensing challenges, such as landfill segmentation. When compared to other lightweight models like ABCNet-E and BA-Net, GLC-Net maintains a distinct advantage in both mIoU and mF1, particularly excelling in vegetation detection. Although these models demonstrate reduced computational cost, they exhibit lower accuracy, underscoring that computational efficiency alone does not guarantee robust performance, especially in heterogeneous remote sensing environments. In comparison to DeSwin-S, GLC-Net demonstrates clear computational efficiency, achieving comparable or even superior segmentation accuracy with fewer parameters and lower FLOPs, further reinforcing the optimal trade-off between model complexity, computational demand, and segmentation performance.

Thus, the lightweight architecture of GLC-Net enables it to rival or surpass more complex models while maintaining a high level of efficiency. This attribute makes GLC-Net particularly suitable for real-time processing in resource-constrained environments, such as mobile platforms or satellite imaging systems, where both high segmentation accuracy and low computational cost are critical. This analysis emphasizes the practical applicability of GLC-Net in complex, heterogeneous remote sensing tasks, positioning it as a viable solution for a wide range of operational scenarios.

2.7 Limitations and future work

While GLC-Net demonstrates strong performance in landfill segmentation through its global-local feature fusion and multi-scale spatial enhancement, several limitations remain. The reliance on pixel-level manual annotations constrains scalability, as labeling is time-consuming and requires expert knowledge. The current datasets, limited to CP and DX, may not fully capture the diversity of urban layouts or waste characteristics in other regions, which could potentially affect generalization. Additionally, the multi-scale deep supervision introduces extra computational overhead during training, and the use of three spectral bands may underexploit the potential of additional multispectral information for finer material discrimination.

Future work will investigate semi-supervised or weakly supervised approaches to reduce annotation dependence and explore knowledge distillation to improve training efficiency. Expanding the dataset to cover diverse geographical regions and integrating additional spectral bands could enhance generalization and material differentiation. Incorporating temporal analysis to monitor landfill dynamics is also planned, enabling more comprehensive applications for urban environmental management.

3 Conclusions

(1) We propose GLC-Net, a global-local collaborative segmentation framework that combines ET-based global context modeling, LB-based local detail enhancement, and MSAE-driven multi-scale spatial aggregation, addressing the challenges of heterogeneous and complex landfill scenes.

(2) Extensive experiments on the CP and DX datasets demonstrate that GLC-Net consistently outperforms CNN- and Transformer-based baselines, achieving 1.5%—3.2% mIoU improvement and su-

perior boundary delineation, validating both the effectiveness and robustness of the proposed framework.

(3) Beyond performance, this work highlights three core contributions: The global-local hybrid framework for feature extraction, the MSAE module for multi-scale context aggregation, and the comprehensive empirical validation across diverse datasets. These contributions collectively provide a robust and generalizable solution for complex remote sensing segmentation tasks.

(4) Future work will focus on reducing annotation dependency, exploring semi-supervised learning, integrating multispectral information, and modeling temporal dynamics to further improve generalization and efficiency.

References

[1] ZHANG C, CHEN Z, LUO L, et al. Mapping urban construction sites in China through geospatial data fusion: Methods and applications[J]. Remote Sensing of Environment, 2024, 315: 114441.

[2] WEI K, DAI J, HONG D, et al. MGFNet: An MLP-dominated gated fusion network for semantic segmentation of high-resolution multi-modal remote sensing images[J]. International Journal of Applied Earth Observation and Geoinformation, 2024, 135: 104241.

[3] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]//Proceedings of Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. Cham: Springer, 2015: 234-241.

[4] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2018: 833-851.

[5] XIE E, WANG W, YU Z, et al. SegFormer: Simple and efficient design for semantic segmentation with Transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 12077-12090.

[6] WANG L, LI R, ZHANG C, et al. UNetFormer: A UNet-like Transformer for efficient semantic segmentation of remote sensing urban scene imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 190: 196-214.

[7] GAO L, LIU H, YANG M, et al. STransFuse: Fusing swin Transformer and convolutional neural network for remote sensing image semantic segmentation[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 10990-11003.

[8] HE X, ZHOU Y, ZHAO J, et al. Swin Transformer embedding UNet for remote sensing image semantic segmentation[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 4408715.

[9] WANG L, LI R, DUAN C, et al. A novel Transformer based semantic segmentation scheme for fine-resolution remote sensing images[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 6506105.

[10] ZHANG Q, YANG Y. ResT: An efficient Transformer for visual recognition[EB/OL]. (2021-05-28). https://arxiv.org/abs/2105.13677.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017. DOI: https://doi.org/10.48550/arXiv.1706.03762.

[12] XIAO T, LIU Y, ZHOU B, et al. Unified perceptual parsing for scene understanding[C]//Proceedings of Computer Vision—ECCV 2018. Cham: Springer, 2018: 432-448.

[13] GHIASI G, LIN T Y, LE Q V. NAS-FPN: Learning scalable feature pyramid architecture for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 7036-7045.

[14] LIN S, HUANG L, LIU X, et al. A construction waste landfill dataset of two districts in Beijing, China from high resolution satellite images[J]. Scientific Data, 2024, 11: 388.

[15] HU M, LI Y, FANG L, et al. A2-FPN: Attention aggregation based feature pyramid network for instance segmentation[C]//Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2021: 15343-15352.

[16] LI R, ZHENG S, ZHANG C, et al. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 181: 84-98.

[17] WANG L, LI R, WANG D, et al. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images[J]. Remote Sensing, 2021, 13(16): 3065.

[18] WU H, HUANG P, ZHANG M, et al. CMTFNet: CNN and multiscale transformer fusion network for remote-sensing image semantic segmentation[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 2004612.

[19] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2017: 618-626.

Acknowledgements This work was supported by the "Fourteenth Five-Year" National Key R&D Program of China (No. 2024YFC3906501) and the New Cornerstone Science Foundation through the XPLORER PRIZE.

Authors

The first author Mr. **WEI Kan** received the B.S. degree from the Faculty of Geoscience and Engineering, Southwest Jiaotong University, Chengdu, China, in 2024. He is currently pursuing his Ph.D. degree at Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include artificial intelligence, multimodal fusion, and image reconstruction.

The corresponding author Prof. **WEN Zongguo** is a tenured professor at School of Environment, Tsinghua University, and Director of the Center for Circular Economy Industry Research. His research focuses on solid waste resource recovery, environmental big data, and intelligent large models. He has received the National Science Fund for Distinguished Young Scholars and the Ten-Thousand Talent Program award, and has published over 220 papers in leading journals including *Nature Sustainability*, *Nature Food*, and *Nature Communications*.

Authors contributions Mr. **WEI Kan** designed the method, conducted the experiments, performed the analysis, and wrote the manuscript. Ms. **LI Ling** wrote the introduction section of the paper. Ms. **LIANG Shilin** wrote the introduction section and supplemented the research background. Prof. **WEN Zongguo** provided supervision, revised the manuscript, and supported the study through funding. All authors commented on the manuscript draft and approved the submmission.

Competing interests The authors declare no competing interests.

(Production Editor: ZHANG Huangqun)

GLC-Net:面向遥感影像高效分割的全局与局部协同网络

魏 刊[1,2], 李 玲[1], 梁诗琳[1], 温宗国[1]

(1.清华大学环境学院,北京100084,中国; 2.中国科学院空天信息创新研究院,北京100190,中国)

摘要:高分辨率空天遥感影像的智能解译是空天信息处理领域的重要研究方向。复杂地表环境,如建筑与拆除(Construction and demolition, C&D)废弃物填埋场等,对遥感影像分割模型的鲁棒性提出了较高要求。传统卷积神经网络(Convolutional neural networks,CNNs)受限于局部感受野,难以捕获全局依赖关系;而基于Transformer的模型虽具备长距离建模能力,却容易忽略细粒度空间结构,导致异质遥感目标分割精度不足。为此,本文提出一种全局与局部协同网络(Global-local collaborative network, GLC-Net),面向空天遥感影像的智能分割任务。该模型融合了高效Transformer模块以建模全局依赖关系,并引入局部增强模块用于细节结构优化。此外,设计了多尺度空间聚合与增强模块(Multi-scale spatial aggregation and enhancement, MSAE)以强化上下文特征表征并抑制背景干扰,同时通过深层监督机制提升多层次语义学习能力。基于两组高分辨率遥感数据集(昌平与大兴)的实验结果表明,GLC-Net在平均交并比(mean intersection over union,mIoU)指标上较现有先进方法提升1.5%~3.2%,并在边界刻画与语义一致性方面表现更优。结果验证了全局-局部协同建模在空天遥感影像智能分割与环境监测中的有效性与潜力。

关键词:遥感影像;深度学习;视觉变换器;填埋场;语义分割