Enhancing MViT for Remote Sensing Segmentation via Adversarial and Frequency-Domain Supervision

Guanlin Li, and Ji Zhao*

***Abstract*—In recent years, visual transformers (ViTs) have demonstrated considerable progress in image classification. However, they continue to face notable challenges in tasks such as high-resolution remote sensing image segmentation and spatiotemporal video understanding. A primary limitation lies in the self-attention mechanism inherent to the Transformer architecture, which tends to emphasize low-frequency information while neglecting high-frequency components. This bias impairs the detection and segmentation of small objects, particularly in the context of remote sensing. To address these limitations, we propose enhancements to the multi-scale visual transformer (MViT) architecture. First, we incorporate a generative adversarial network framework based on Wasserstein GAN with gradient penalty (WGAN-GP) to improve the segmentation of small objects. This approach introduces a discriminator loss that encourages the model to focus more effectively on small-scale features. Second, we design a loss function that leverages frequency-domain image characteristics to mitigate the loss of fine-grained details during training. Furthermore, we integrate the optimized MViT architecture with the standard dense prediction framework, SegFormer, to enhance segmentation accuracy on complex remote sensing images. Experimental results demonstrate that our proposed model, FGAMViT, significantly improves the ability to capture multi-scale and complex features while maintaining a favourable trade-off between computational efficiency and segmentation performance. These enhancements offer a more robust solution for applying Transformer-based models to remote sensing image segmentation tasks.**

***Index Terms*—Image Segmentation, Deep Learning, Vision Transformer, Self-Attention**

I. Introduction

DESIGNING effective architectures for remote sensing image segmentation has long been a challenging task due to the high spatial resolution, complex textures, and varying object scales present in such data [1, 2]. Traditional convolutional neural network (CNN)-based models, such as ResNet [3, 4] and U-Net [5–7], have been widely adopted due to their simplicity and computational efficiency. However, recent advances in Vision Transformers (ViTs) have demonstrated remarkable performance across diverse vision tasks, challenging the long-standing dominance of Convolutional Neural Networks (CNNs) [8–10]. The original ViT architecture, although designed for image classification, has since been adapted and extended to meet the specific demands of various visual tasks, including dense prediction and temporal modelling.

Despite their success in image classification, ViTs face significant hurdles in high-resolution remote sensing image segmentation and spatiotemporal video understanding [11–13]. Remote sensing images often contain intricate land cover patterns, multiple object categories, and diverse spatial contexts, all captured at very high resolutions [14]. These factors introduce substantial computational and memory burdens, primarily because the self-attention mechanism in standard Transformers scales quadratically with the input resolution. As a result, applying vanilla ViT models directly to large-scale inputs becomes infeasible without modification. To address these issues, a variety of architectural strategies have been proposed, among which two prominent directions have emerged: window-based attention and pooled multi-scale attention. The first strategy involves window-based attention mechanisms, exemplified by the Swin Transformer [15, 16]. Swin Transformer introduces a hierarchical structure in which self-attention is computed within non-overlapping local windows. To expand the receptive field and capture long-range dependencies, it employs a shifted window strategy across layers. This hierarchical design reduces the computational cost while preserving the model's ability to learn both local and contextual information. Consequently, the Swin Transformer has demonstrated strong performance in remote sensing tasks, including land cover classification, object detection, and semantic segmentation. However, its localized attention design can still limit its capacity to capture truly global contextual cues, mainly when the semantic meaning of an object is distributed across widely separated regions in high-resolution scenes.

An alternative and promising direction is pooled or hierarchical attention, as demonstrated by the Multiscale Vision Transformer (MViT) [17, 18]. MViT introduces a multiscale tokenization and attention mechanism that progressively downsamples and aggregates feature representations across layers. This hierarchical fusion of spatial information allows the model to capture both fine-grained local details and high-level global semantics. MViT is particularly well-suited to processing remote sensing data, as it can simultaneously handle objects of varying sizes and leverage contextual dependencies at different spatial resolutions. By integrating multiscale features into the attention mechanism, MViT has proven effective in fine-grained segmentation and small object detection, outperforming traditional CNNs in several benchmarks.

In this study, we enhance the MViT architecture better to

Manuscript received May 7, 2025; revised Jul 4, 2025. The research work was supported by a scientific research project fund from the Liaoning Provincial Department of Education,and key project of Liaoning Provincial Department of Education(LJKZZ2022043)

Guanlin Li is a Postgraduate of University of Science and Technology Liaoning, Anshan, Liaoning, China (e-mail:LGLin_2000@outlook.com).

Ji Zhao* is a Professor of University of Science and Technology Liaoning, Anshan, Liaoning, China (corresponding author to provide phone: +086-139-9808-6167; e-mail:zhaoji_1974@126.com).

meet the challenges of remote sensing image segmentation. The proposed improvements are as follows:

1) Integration of a Generative Adversarial Network Framework (WGAN-GP) [19]: Conventional loss functions such as L1, L2, or Binary Cross Entropy (BCE) often fail to capture the distributional characteristics of small objects, especially when such targets are sparse or occupy only a small fraction of the image. These losses tend to focus disproportionately on significant regions, such as the background, leading to suboptimal segmentation performance. To address this, we incorporate a Wasserstein GAN with gradient penalty (WGAN-GP) as an auxiliary supervision mechanism. The discriminator learns to distinguish between the ground truth and the predicted segmentation maps, encouraging the generator to produce more realistic and structurally coherent outputs. This adversarial learning framework provides smoother gradient updates and promotes finer attention to small-scale regions, thereby enhancing the model's ability to recover small target structures.

2) Frequency-Domain Loss Function: Remote sensing images often exhibit a spectral imbalance, where low-frequency components dominate the spatial distribution, and high-frequency components—such as edges, boundaries, and small objects—are sparse yet critical. To counteract the model's inherent bias toward learning low-frequency content, we propose a frequency-aware loss based on the 2D Fourier transform. This loss penalizes discrepancies in the spectral domain, with a focus on preserving high-frequency information relevant to object boundaries and delicate textures. By adaptively down-weighting the loss contribution of easy-to-synthesise (low-frequency) components and emphasising the reconstruction of hard-to-synthesise (high-frequency) ones, the proposed loss acts as a complement to traditional pixel-wise spatial losses. It effectively preserves the structural details essential for the precise segmentation of small and complex targets.

3) Integration with SegFormer: We incorporate our optimized MViT into the SegFormer dense prediction framework [20], further enhancing its capability to handle complex segmentation tasks in remote sensing images. Experimental results demonstrate that these optimizations enhance the model's ability to handle multi-scale and intricate features, resulting in substantial performance gains in remote sensing image segmentation. By combining the improved MViT with SegFormer, our approach achieves a strong balance between computational efficiency and segmentation accuracy, offering a powerful solution for Transformer-based models in remote sensing applications.

II. Related Work

A. From Convolutional Networks to Vision Transformers in Remote Sensing Image Segmentation

Convolutional Neural Networks (CNNs) have long been the foundation of remote sensing image segmentation due to their ability to extract spatial features through local convolution operations in a hierarchical manner [21, 22]. Architectures such as U-Net [23], DeepLab [24], and ResNet-based backbones [25] have been widely adopted in this domain, primarily due to their efficiency, simplicity, and firm performance on various pixel-wise prediction tasks. Their capacity to model local patterns and semantic hierarchies makes them effective for extracting texture, edges, and mid-level features. However, CNNs inherently suffer from limited receptive fields and inductive biases such as translation invariance, which constrain their ability to capture long-range dependencies and global contextual relationships—critical for understanding complex and large-scale remote sensing scenes.

To overcome these limitations, researchers have increasingly turned to Transformer-based models, initially developed for natural language processing [26], and adapted them to vision tasks. Vision Transformers (ViTs) [27] has introduced a paradigm shift by replacing convolutions with self-attention mechanisms, enabling models to capture global dependencies across an entire image. In the ViT architecture, an image is divided into fixed-size patches, each of which is linearly embedded and treated as a token in a sequence. This sequence is then processed by Transformer layers that learn relationships between all patches, regardless of their spatial proximity. This approach allows ViTs to model long-range interactions and complex spatial dependencies more effectively than CNNs.

Despite their success in image classification, the application of ViTs to high-resolution remote sensing images introduces new challenges. A key issue lies in the computational burden of the self-attention mechanism, whose complexity scales quadratically with the number of input tokens [28, 29]. In high-resolution tasks, where the number of patches can be huge, this results in substantial memory and processing demands. Moreover, ViTs lack the inductive biases of CNNs, which can hinder learning in data-scarce regimes typical of remote sensing applications.

To address these drawbacks, several hybrid or hierarchical Transformer architectures have been proposed. For instance, the Swin Transformer employs a local window-based self-attention strategy with hierarchical feature aggregation, significantly reducing the computational cost while preserving the ability to capture global patterns through shifted windows [30]. Similarly, models like the Pyramid Vision Transformer (PVT) and SegFormer integrate CNN-like multi-scale feature extraction with Transformer-based global context modelling, offering a favourable balance between efficiency and representational power [31]. These developments have significantly advanced the state-of-the-art in remote sensing image segmentation, enabling Transformer-based models to handle large-scale and complex scenes more effectively.

B. Multi-Scale Vision Transformer (MViT)

The Multi-Scale Vision Transformer (MViT) was proposed to overcome the inherent limitations of the original Vision Transformer (ViT), specifically its inability to process multi-scale information and its high computational complexity. While ViT treats image patches uniformly and lacks an explicit mechanism for hierarchical feature extraction, MViT introduces a progressive token pooling and multi-scale attention strategy, allowing the network to learn visual representations at different resolutions. This hierarchical design enables MViT to capture both fine-grained local details and coarse global context, making it well-suited for visual tasks involving objects of varying sizes and complex spatial relationships.

In the MViT architecture, tokens are progressively downsampled through pooling operations across layers, which reduces the computational burden while preserving essential spatial structure. Simultaneously, multi-scale attention modules enable information exchange across different resolutions, thereby enhancing the model's capability to learn both high-level semantics and low-level textures. These properties are particularly advantageous for remote sensing image segmentation, where scenes often contain small, scattered objects embedded within large-scale and heterogeneous backgrounds.

Despite its strengths, MViT still faces challenges when applied to high-resolution remote sensing images. The primary issue lies in accurately segmenting small objects and preserving fine-grained details, which can be lost during the progressive pooling and token downsampling operations. Additionally, while the model captures multi-scale information, it does not explicitly emphasise difficult-to-learn components, such as high-frequency textures or small structures, which often carry critical semantic information in remote sensing tasks.

C. SegFormer

SegFormer is a Transformer-based architecture designed specifically for image segmentation tasks. It effectively combines the strengths of models like PVT, which excel at multi-scale feature extraction, with the global dependency modelling capabilities of Transformers. SegFormer employs an efficient self-attention mechanism and a simplified decoder, achieving an optimal balance between computational efficiency and segmentation accuracy. This design enables the model to produce high-quality segmentation results while maintaining a low computational cost.SegFormer has demonstrated outstanding performance across a range of semantic segmentation tasks, highlighting its versatility and robustness in handling complex scenarios. Its ability to process intricate visual details and diverse image structures makes it especially suited for challenging applications, such as remote sensing image segmentation. With its efficiency and scalability, SegFormer is a powerful solution for tackling the complexities of segmenting high-resolution, multi-scale remote sensing images.

III. FGAMViT Architecture

As illustrated in Figure 1, the FGAMViT architecture consists of two primary components: a generator network and a discriminator network. The generator integrates the Multi-Scale Vision Transformer (MViT) with the SegFormer framework to perform remote sensing image segmentation (Figure 1.A), while the discriminator network evaluates the authenticity of the segmentation output by distinguishing between authentic segmentation masks and those produced by the generator (Figure 1.E). The two networks are trained in an adversarial manner: the generator minimises the Wasserstein distance [32] between the real and generated segmentation distributions. At the same time, the discriminator seeks to maximise this distance, thereby providing adversarial feedback that guides the generator toward producing more realistic and accurate segmentation maps.

A. Generative Component

The generator, based on a Transformer backbone, constitutes the core of the FGAMViT architecture [26]. Transformers have demonstrated strong capabilities in modelling long-range dependencies within sequences, which has contributed to their widespread success across a range of vision tasks [28]. However, the conventional multi-head self-attention mechanism used in Transformers incurs significant computational and memory costs. The complexity scales quadratically with input size, making the architecture less feasible for large-scale, high-resolution remote sensing imagery.

To address this limitation, FGAMViT incorporates pooling attention mechanisms—a design that restricts the attention range to reduce computational burden while preserving essential local and global dependencies. This enables the efficient modelling of complex spatial patterns in remote sensing images without a substantial trade-off in performance.

As depicted in Figure 1.A, the overall structure of the generator follows the SegFormer design paradigm, which includes a hierarchical encoder-decoder architecture. The encoder comprises four MViT blocks, each tasked with nonlinearly mapping the input remote-sensing image into feature representations of varying spatial resolutions. Each MViT block integrates three critical components:

1) Pooling Self-Attention Module: a modified attention mechanism that balances efficiency and accuracy;
2) Mix Feed-Forward Network (Mix-FFN): responsible for feature transformation;
3) Overlap Patch Merging: enabling hierarchical feature integration.

The Pooling Self-Attention module, as illustrated in Figure 1.D, is further augmented with residual pooling connections. After pooling operations, residual connections are introduced to combine the pooled output with the original feature map. This fusion preserves high-frequency and fine-grained details that are typically lost during downsampling, thereby enriching the representation passed to subsequent layers. This mechanism not only retains essential spatial cues but also enhances the model's ability to delineate object boundaries and recognize subtle texture variations—key requirements for remote sensing segmentation tasks.

The process begins with input normalization (Equation 1), where X represents the raw input features and X' denotes the normalized features. This step ensures that the input is appropriately scaled for the subsequent operations. Following this, a linear transformation is applied to the normalised features to produce the query (Q), key (K), and value (V) vectors, as shown in Equation (2). These transformations are performed using specific weight matrices W_Q, W_K, and W_V for each respective vector. The goal is to enable the model to capture various aspects of the input features, which are crucial for computing the attention scores. To reduce computational complexity, the query, key, and value vectors are combined in the next step, as shown in Equation (3). Pooling these vectors helps decrease the overall computational load, especially when working with high-resolution data, while still retaining enough information for meaningful attention calculations.

$$\hat{X} = \mathrm{Norm}(X) \tag{1}$$

$$Q = W_Q \hat{X}, \quad K = W_K \hat{X}, \quad V = W_V \hat{X} \tag{2}$$

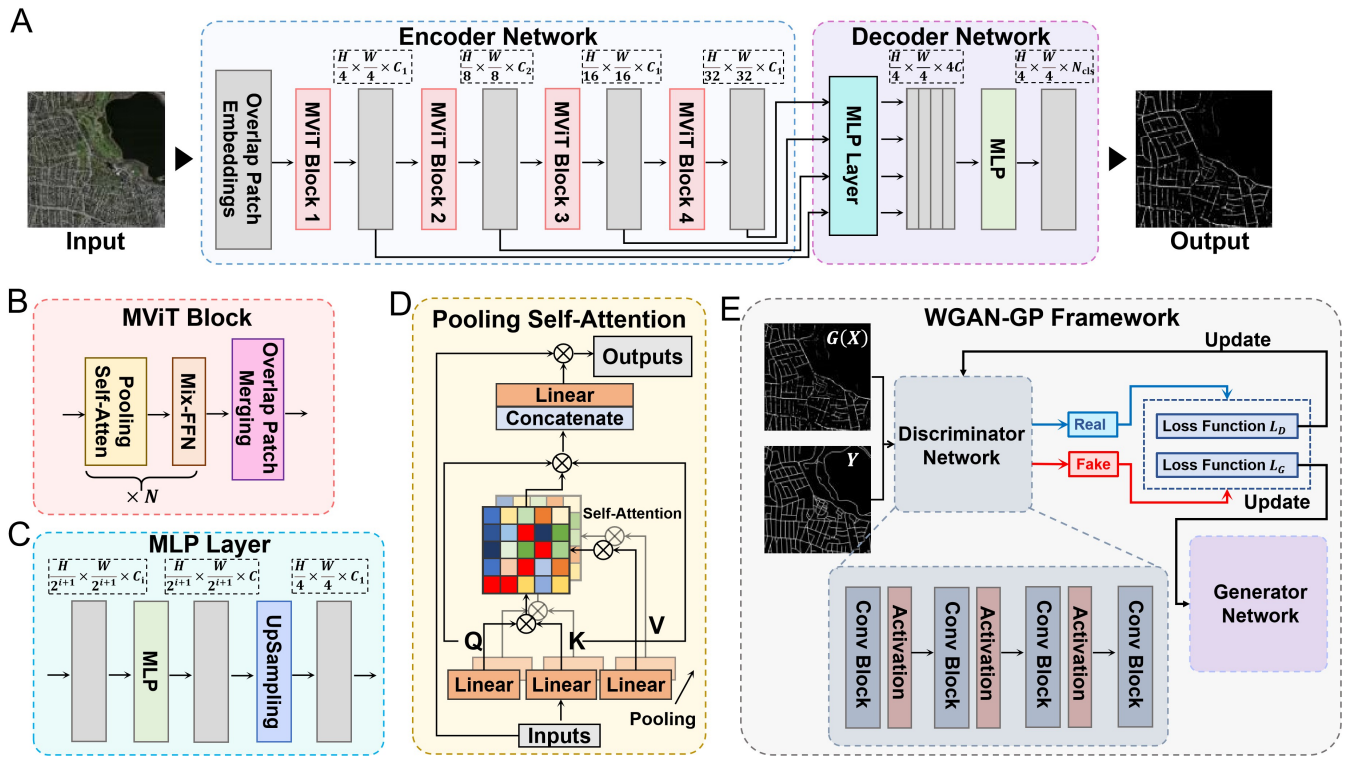

Fig. 1. **Overall architecture of FGAMViT.**

$$Q_{\text{pool}} = \text{Pool}(Q), \quad K_{\text{pool}} = \text{Pool}(K), \quad V_{\text{pool}} = \text{Pool}(V) \tag{3}$$

Next, relative positional embeddings are computed and integrated into the attention scores Equation (4). The embeddings $E(i,j)$ and $F(i,j)$ represent the relative positions of elements i and j in the input sequence. These embeddings enable the model to be more robust to positional shifts and enhance its spatial awareness, making it more adaptable to changes in object locations within the image. The attention weights are then determined using the Softmax function, which normalizes the attention scores across all elements. These normalised weights are used to weight the value vectors, which are then summed to produce the self-attention output (Equation 5). This allows the model to focus on the most relevant information based on the computed attention. To retain important features from the original input, a residual connection is introduced by adding the original input features (before pooling) to the self-attention output (after pooling), as shown in Equation (6). This residual link helps preserve crucial information lost due to pooling, ensuring better feature retention and aiding in maintaining the integrity of the original input. Finally, the output from the residual connection undergoes a linear transformation, as specified in Equation (7), using the transformation matrix W_O, which generates the model's final output. This output is then passed to subsequent layers or operations.

$$A_{ij} = \frac{Q_{\text{pool},i} K_{\text{pool},j}^{\top}}{\sqrt{d}} + E(i,j) + F(i,j) \tag{4}$$

$$\alpha_{ij} = \text{Softmax}(A_{ij}) \tag{5}$$

$$Z = \sum_{j} \alpha_{ij} V_{\text{pool},j} \tag{6}$$

$$\text{Output} = W_O Z_{\text{res}} \tag{7}$$

By combining the standard self-attention mechanism with pooling operations, residual connections, and relative positional embeddings, this process significantly improves the model's ability to capture and maintain positional information. It also ensures that detailed features are preserved while reducing computational complexity, making the model more efficient and effective, particularly for tasks such as remote sensing image segmentation, where both accuracy and computational efficiency are crucial.

This architectural refinement enables the generator to strike a balance between computational efficiency and segmentation accuracy, allowing it to process high-resolution imagery effectively. By minimizing the degradation of critical spatial features, the generator can more precisely capture complex structures in remote sensing scenes, ultimately contributing to superior segmentation performance.

B. Discriminative Component

To more effectively quantify the discrepancy between the predicted segmentation map and the corresponding ground truth, we design a discriminator network inspired by the Patch-GAN architecture [33]. This design enables the network to focus on local-level consistency rather than global realism, which is more suitable for segmentation tasks. The discriminator maps the input—either real or generated segmentation results—to a high-dimensional logit representation, providing spatially localized feedback to guide the generator in refining segmentation quality.

As illustrated in Figure 1.E, the discriminator receives two types of inputs: (1) a concatenation of the original remote sensing image and the generator-produced segmentation output used to compute pseudo logits and (2) a concatenation of the original input image and the corresponding ground truth segmentation map, used to compute real logits. These

two inputs are passed through the same network structure to evaluate their authenticity.

The architecture of the discriminator consists of a series of convolutional blocks interleaved with pooling and activation operations. Each convolutional block includes A 2D convolutional layer for feature extraction, a 2×2 max pooling operation for spatial downsampling, and a ReLU activation function for non-linearity.

Initially, both inputs are processed by a convolutional layer that projects them into a feature space of size $128 \times 128 \times 64$. This is followed by repeated downsampling and feature refinement stages. Specifically, the input is progressively reduced through four such convolution-pooling-activation cycles, ultimately resulting in feature maps of dimension $16 \times 16 \times 64$. Finally, a 1×1 2D convolution is applied to compress the feature map into a logits tensor of size $16 \times 16 \times 1$, where each value represents the discriminator's confidence in the authenticity of a corresponding spatial patch in the segmentation map.

During adversarial training, the discriminator serves a dual role. The pseudo logits derived from generated segmentation outputs are used to update the generator network via the adversarial loss, encouraging it to produce outputs that are increasingly indistinguishable from real segmentations. Simultaneously, the discriminator is updated using both the real and fake logits, thereby enhancing its ability to distinguish authentic segmentation maps from synthetic ones. This adversarial interaction strengthens the generator's ability to produce fine-grained, spatially consistent segmentations and enhances the discriminator's sensitivity to structural inconsistencies—critical for high-resolution remote sensing image analysis.

C. Training Process of FGAMViT

The loss function used for training FGAMViT consists of two components: the generator loss L_G and the discriminator loss L_D. The generator loss combines pixel-level reconstruction, adversarial supervision, and frequency-domain alignment. It is defined as:

$$L_G = \mathbb{E}_{Y,X}\left[(G(X) - Y)^2\right] + \alpha \mathbb{E}_X\left[D(G(X), X)\right] + \beta L_F \tag{8}$$

The first term is the pixel-wise mean square error (MSE) between the generated segmentation output, $G(X)$, and the ground truth label, Y, which ensures spatial accuracy in segmentation. The second term, weighted by the hyperparameter α, represents the adversarial loss contributed by the discriminator's evaluation of the generated segmentation. The third term, L_F, represents the frequency-domain loss, which is described as follows.

a) Frequency-Domain Loss L_F: To enhance the preservation of fine-grained details and high-frequency components, which are often underrepresented in remote sensing segmentation, we introduce a loss term based on the 2D Discrete Fourier Transform (DFT). Let $\mathcal{F}(\cdot)$ denote the 2D DFT. We define:

$$\hat{G}(u, v) = \mathcal{F}(G(X)), \quad \hat{Y}(u, v) = \mathcal{F}(Y) \tag{9}$$

Where $\hat{G}(u, v)$ and $\hat{Y}(u, v)$ are the frequency representations of the generated and real segmentation maps, respectively. The frequency loss is then calculated as:

$$L_F = \frac{1}{HW} \sum_{u=1}^{H} \sum_{v=1}^{W} w(u, v) \cdot \left| \hat{G}(u, v) - \hat{Y}(u, v) \right|^2 \tag{10}$$

with a frequency weighting function defined by:

$$w(u, v) = \log(1 + u^2 + v^2) \tag{11}$$

This weighting emphasizes higher-frequency components (e.g., edges, contours), promoting sharper predictions while maintaining training stability.

b) Discriminator Loss L_D: The discriminator is trained to distinguish between real and generated segmentations. Its loss function follows the WGAN-GP formulation and is defined as:

$$\begin{aligned} L_D = {} & \mathbb{E}_X\left[D(G(X), X)\right] - \mathbb{E}_{Y,X}\left[D(Y, X)\right] \\ & + \gamma \left(\mathbb{E}\left[\left\| \nabla_{\widetilde{X}} D(\widetilde{X}) \right\|_2 \right] - 1 \right) \end{aligned} \tag{12}$$

Where $\widetilde{X} = \varepsilon X + (1 - \varepsilon) G(X)$ is a linear interpolation between the real and generated inputs, and the third term enforces the gradient penalty to stabilize adversarial training, following [19]. Here, β is the weight of the penalty term, and ε is sampled from a uniform distribution, fixed to 0.5 in our experiments.

c) Training Details: FGAMViT is trained for 50 epochs with a batch size of 4. The generator and discriminator use learning rates of $\epsilon = 2 \times 10^{-4}$ and $\epsilon = 1 \times 10^{-4}$, respectively. The generator is optimized using AdamW, while the discriminator uses RMSprop. The coefficients α, β and γ are set to -1×10^{-3}, -5×10^{-4} and 10, respectively. All experiments are accelerated using an NVIDIA RTX 3090 graphics processing unit (GPU).

IV. Experimental Setting

A. Datasets

To train and evaluate the FGAMViT architecture, we utilize four publicly available remote sensing datasets commonly used in building segmentation research: the INRIA Aerial Image Labeling Dataset (IAIL) [34], SpaceNet [35], DeepGlobe Building Extraction Dataset (DBE) [36], and WHU Building Dataset (WHUB) [37]. These datasets provide high-resolution imagery with pixel-level annotations, enabling robust training and benchmarking of segmentation models.

a) INRIA Aerial Image Labeling Dataset (IAIL): The IAIL dataset consists of 180 high-resolution aerial images, each with a resolution of 5120×5120 pixels and a spatial resolution of 0.3 meters. The dataset encompasses urban and suburban scenes from cities including Austin, Chicago, and Vienna. Each pixel is labelled as either "building" or "non-building," making it suitable for binary segmentation tasks. For model training, we divide the dataset into 60% for training, 20% for validation, and 20% for testing. All images are tiled into patches of 512×512 pixels with 50% overlap and are normalised to have a zero mean and unit variance.

b) SpaceNet Dataset: The SpaceNet dataset provides high-resolution satellite images (0.3–0.5m GSD) from several global cities with dense, diverse urban structures. Each image includes pixel-wise annotations of the building footprint. We utilise the SpaceNet Building Detection Challenge subset and apply a 70%-15%-15% split for training, validation, and testing, respectively. Images are resized to 512×512 patches and augmented using random rotation, horizontal flipping, and contrast normalization to enhance generalization.

c) DeepGlobe Building Extraction Dataset (DBE): The DBE dataset includes 650×650 pixel satellite image patches with 0.5m resolution, covering urban, suburban, and rural areas. Each patch is annotated with building contours. We partition the dataset into 70% training, 15% validation, and 15% testing sets. Preprocessing includes resizing to 512×512, histogram equalization, and data augmentation (random cropping, flipping, and brightness adjustment).

d) WHU Building Dataset (WHUB): The WHUB dataset comprises 30,000 aerial images spanning 450 square kilometres across various urban and rural areas in China. The spatial resolution ranges from 0.3 m to 0.5 m. Each pixel is labelled as either "building" or "non-building." For this dataset, we follow the standard protocol and use 80% of the data for training, 10% for validation, and 10% for testing. Each image is divided into 512×512 tiles, with normalisation and shadow removal preprocessing applied.

e) Preprocessing and Augmentation: All datasets are unified into a common resolution of 512×512 pixels during preprocessing. We perform standardisation (zero mean, unit variance) and apply consistent data augmentation strategies, including random flipping, rotation (by 90° increments), brightness and contrast adjustments, and Gaussian noise addition. This improves model robustness to diverse imaging conditions.

B. Evaluation Metrics

To comprehensively assess the performance of FGAMViT on building segmentation tasks, we employ a set of widely used quantitative metrics that evaluate both segmentation accuracy and computational efficiency.

a) Mean Intersection over Union (mIoU): mIoU is the primary metric for segmentation quality, defined as the ratio of the intersection to the union of the predicted and ground truth masks:

$$\text{IoU} = \frac{TP}{TP + FP + FN}, \quad \text{mIoU} = \frac{1}{C} \sum_{i=1}^{C} \text{IoU}_i \tag{13}$$

Where C is the number of classes (here $C = 2$ for building and non-building), and TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

b) F1 Score: The F1 score balances precision and recall, which is particularly important for imbalanced datasets with sparse building regions:

$$\begin{aligned} \text{F1} &= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \\ \text{Precision} &= \frac{TP}{TP + FP}, \\ \text{Recall} &= \frac{TP}{TP + FN} \end{aligned} \tag{14}$$

c) Dice Coefficient: The Dice coefficient measures the overlap between prediction and ground truth and is particularly effective for evaluating segmentation shapes:

$$\text{Dice} = \frac{2TP}{2TP + FP + FN} \tag{15}$$

d) Pixel Accuracy: Pixel accuracy is defined as the ratio of correctly classified pixels to the total number of pixels:

$$\text{Pixel Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

Although intuitive, this metric may be biased in scenes where building pixels occupy a small fraction of the total image area.

e) Floating Point Operations (FLOPs): FLOPs indicate the number of floating point operations required during inference. While not directly related to segmentation accuracy, lower FLOPs indicate better computational efficiency, which is crucial for resource-constrained environments.

f) Frames Per Second (FPS): FPS measures real-time inference speed, defined as the number of image frames processed per second. High frame rates (FPS) are vital for time-sensitive applications, such as drone surveillance and disaster monitoring.

Among the above, mIoU, F1 Score, Dice Coefficient, and Pixel Accuracy range from 0 to 1 (reported as percentages), with higher values indicating better segmentation performance. FLOPs should be minimised to reduce computational cost, while FPS should be maximised to ensure real-time performance. These complementary metrics provide a balanced evaluation of segmentation accuracy and model efficiency.

C. Baseline Models

For remote sensing image segmentation tasks, convolutional neural network (CNN) architectures such as U-Net, DenseUNet [38], ResUNet [39], and DeepLabV3+ [40] offer distinct advantages in extracting building footprints and handling complex background environments. Additionally, Transformer-based models like SegFormer also provide significant benefits. The following provides a brief overview of the baseline models evaluated in this study:

a) U-Net: U-Net is a widely adopted encoder-decoder architecture developed initially for biomedical image segmentation. It has since become a foundational model in remote sensing applications due to its ability to capture spatial hierarchies and recover fine details. The skip connections between the encoder and decoder enable precise localisation, making U-Net effective for segmenting buildings with well-defined boundaries, even in heterogeneous scenes.

b) U-Net++: U-Net++ is an enhanced variant of the original U-Net architecture, designed to improve the accuracy and efficiency of image segmentation tasks through a nested and dense skip connection structure. It introduces intermediate convolutional layers between encoder and decoder pathways to reduce the semantic gap and improve feature fusion. This architectural refinement enables U-Net++ to capture multiscale contextual information better and enhance boundary delineation. In remote sensing applications, U-Net++ is particularly effective for segmenting buildings with complex shapes and varying scales, offering improved performance over traditional encoder-decoder frameworks in challenging urban environments.

c) PSPNet: Pyramid Scene Parsing Network (PSPNet) enhances semantic segmentation by capturing global contextual information through its pyramid pooling module. This module aggregates features at multiple spatial scales, allowing the network to understand both coarse and fine spatial patterns. PSPNet is particularly effective in remote sensing scenarios, where buildings and landscapes exhibit diverse sizes and spatial distributions. Its ability to integrate multi-scale context significantly improves segmentation performance in complex urban environments, especially in scenes with occlusions, shadows, and cluttered structures.

d) DenseUNet: DenseUNet incorporates dense connectivity into the U-Net structure, where each layer receives input from all preceding layers within the block. This design promotes feature reuse and strengthens gradient flow, improving the segmentation of small objects and fine-building structures. Its enhanced representation capability makes it suitable for remote sensing imagery with dense and diverse building distributions.

e) ResUNet: ResUNet integrates residual blocks into the U-Net architecture, enabling deeper network structures without degradation. The residual connections help preserve spatial information during encoding and facilitate better gradient propagation during training. ResUNet has been proven effective in segmenting buildings in high-resolution aerial and satellite imagery, especially in scenarios involving shadow occlusion and varying architectural forms.

f) DeepLabV3+: DeepLabV3+ extends the DeepLab family by combining atrous spatial pyramid pooling (ASPP) with an encoder-decoder structure. It efficiently captures multi-scale contextual information and refines object boundaries, which is critical for segmenting buildings at varying scales. The use of dilated convolutions enhances the receptive field without loss of resolution, making it robust in complex urban landscapes.

g) SegFormer: SegFormer is a Transformer-based semantic segmentation model that combines lightweight hierarchical encoding with a simple MLP-based decoder. It leverages self-attention to model global dependencies, making it particularly effective in capturing large-scale spatial relationships and accurately segmenting buildings even under heavy occlusion or low-contrast conditions.

Each of these models presents unique strengths suited to specific challenges in remote sensing image segmentation. U-Net and its variants (DenseUNet, ResUNet) excel in preserving spatial detail and adapting to diverse urban structures. DeepLabV3+ strikes a balance between efficiency and accuracy through multi-scale context aggregation, while SegFormer offers superior performance in capturing global patterns. Together, these baselines provide a comprehensive foundation for evaluating the proposed FGAMViT architecture.

V. Experimental and Analysis

A. Visualized Results

Figure 2 presents the qualitative segmentation results of FGAMViT across four benchmark remote sensing datasets. Specifically, Figure 2.A illustrates the results on the IAIL dataset, Figure 2.B on the SpaceNet dataset, Figure 2.C on the DeepGlobe Building Extraction dataset, and Figure 2.D on the WHU Building dataset.

For each dataset, the subfigures from left to right show the input remote sensing image, the predicted segmentation map generated by FGAMViT, and the corresponding ground truth annotation. The visual comparisons demonstrate that FGAMViT effectively captures fine-grained details and complex structures, achieving high correspondence with the annotated ground truths across diverse urban and rural scenes.

To highlight the performance advantages of the FGAMViT model, Figure 3 presents a normalized comparative visualization of evaluation metrics (excluding FLOPs) between various baseline models and FGAMViT across different datasets, with FGAMViT's performance standardized to 100 as the benchmark. This intuitive comparison approach clearly demonstrates the relative performance differences among models across multiple dimensions. Such analytical methodology not only facilitates a deeper understanding of model characteristics, but also provides essential guidance for subsequent model optimization and selection.

B. Model Comparison

Experiment 1 evaluates the proposed FGAMViT model against a set of representative baseline segmentation models, including U-Net, U-Net++, PSPNet, DenseUNet, ResUNet, DeepLabV3+, and SegFormer, across four benchmark remote sensing datasets. Table I presents the comparative results, with the best and second-best scores highlighted in bold and underlined, respectively. FGAMViT consistently ranks first or second across all key performance metrics, confirming its robustness and effectiveness in building segmentation.

a) IAIL Dataset: FGAMViT achieves the highest mIoU (41.47%), F1 Score (86.72%), and Pixel Accuracy (84.43%), outperforming the second-best model by margins of 2.71%, 8.74%, and 7.76%, respectively. It ranks second in Dice Coefficient (77.00%) and FLOPs, trailing the top model by 2.12% and 2.49%. FPS reaches 15.9, slightly below the highest rate of 17.5, reflecting a trade-off for enhanced segmentation accuracy.

b) SpaceNet Dataset: FGAMViT leads in mIoU (37.02%), F1 Score (76.51%), and Pixel Accuracy (74.83%), surpassing the runner-up by 0.73%, 9.57%, and 6.59%, respectively. It ranks second in Dice Coefficient (68.37%) and FLOPs, with minimal differences of 2.24% and 2.91%. Its FPS (14.0) is competitive, with only a 10.3% gap compared to the fastest model.

c) DeepGlobe Building Extraction Dataset: On this dataset, FGAMViT again secures the highest mIoU (40.18%), F1 Score (82.01%), and Pixel Accuracy (80.13%), exceeding the second-best by 1.04%, 10.27%, and 7.12%, respectively. It ranks second in Dice Coefficient (72.28%) and FLOPs, with small gaps of 2.80% and 2.53%. FPS (14.9) remains within an acceptable range for high-accuracy applications.

d) WHU Building Dataset: FGAMViT demonstrates the highest F1 Score (80.67%), Dice Coefficient (74.64%), and Pixel Accuracy (80.18%), outperforming the second-best model by 6.81%, 2.49%, and 6.73%, respectively. It ranks second in mIoU (38.78%) and FLOPs, with differences of 1.42% and 2.54%. The FPS (16.4) trails the fastest baseline (18.2) by 9.9%.

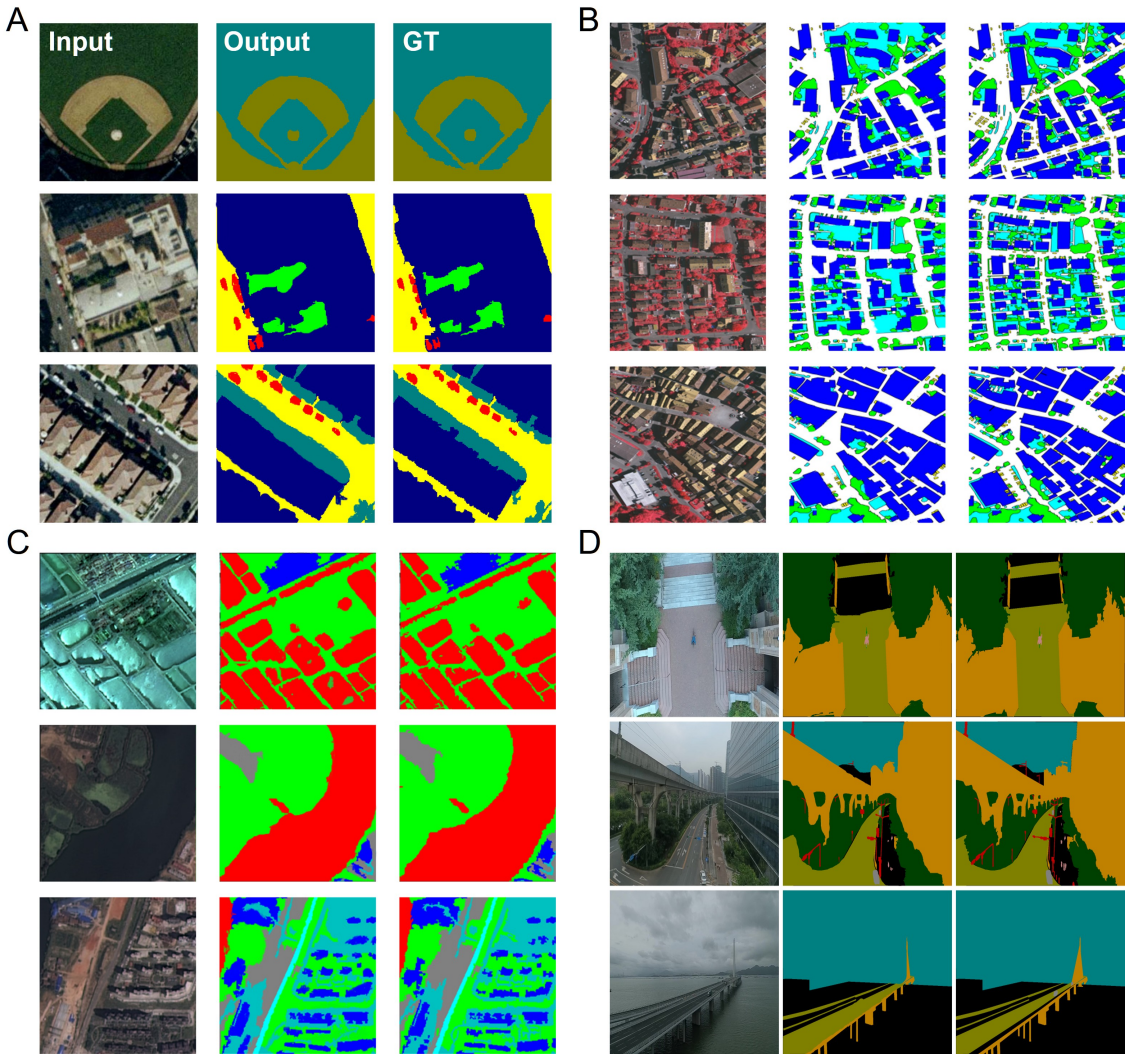

Fig. 2. **Segmentation results of FGAMViT.**

e) Metric Analysis: We adopt five core metrics to assess model performance:

Mean Intersection over Union (mIoU): mIoU measures the average overlap between the predicted and ground-truth regions. FGAMViT consistently achieves the highest or second-highest mIoU, reflecting its ability to maintain structural consistency across diverse urban layouts.

F1 Score: As the harmonic mean of precision and recall, the F1 Score is crucial for evaluating class-imbalanced datasets. FGAMViT's high F1 scores across all datasets highlight its reliability in detecting buildings with minimal false positives and negatives.

Pixel Accuracy: While this metric can favour dominant background classes, FGAMViT's consistently high pixel accuracy confirms its overall prediction quality. When combined with mIoU and F1 Score, it offers a comprehensive view of performance.

Dice Coefficient: Although FGAMViT slightly trails SegFormer in Dice Coefficient on some datasets, it remains competitive. The marginal differences are mainly due to SegFormer's simplified decoder, which favours sharp boundary delineation. FGAMViT compensates with more context-aware, globally consistent predictions.

FLOPs and FPS: FGAMViT exhibits moderate computational complexity. While its FPS (ranging from 14.0 to 16.4) is marginally lower than the most lightweight models, its FLOPs remain within an efficient range. The model prioritizes segmentation quality, making the trade-off appropriate for scenarios where accuracy is paramount.

Although FGAMViT does not achieve the highest frame rate, it consistently delivers superior or near-optimal performance across all critical segmentation metrics. Its strong results in mIoU, F1 Score, and Pixel Accuracy make it well-suited for high-resolution remote sensing applications requiring detailed and reliable building extraction. FGAMViT offers a balanced solution that effectively combines segmentation fidelity with computational efficiency.

C. Impact of Discriminator

Experiment 2: Evaluating the Contribution of the Discriminator Network to FGAMViT Performance

Building upon the high segmentation accuracy demonstrated by FGAMViT in Experiment 1, Experiment 3 explores the role of the discriminator component in enhancing the model's predictive quality. Specifically, this experiment investigates how the integration of a WGAN-GP-based discriminator affects the model's ability to refine segmentation boundaries and recover small-scale structures. The WHUB

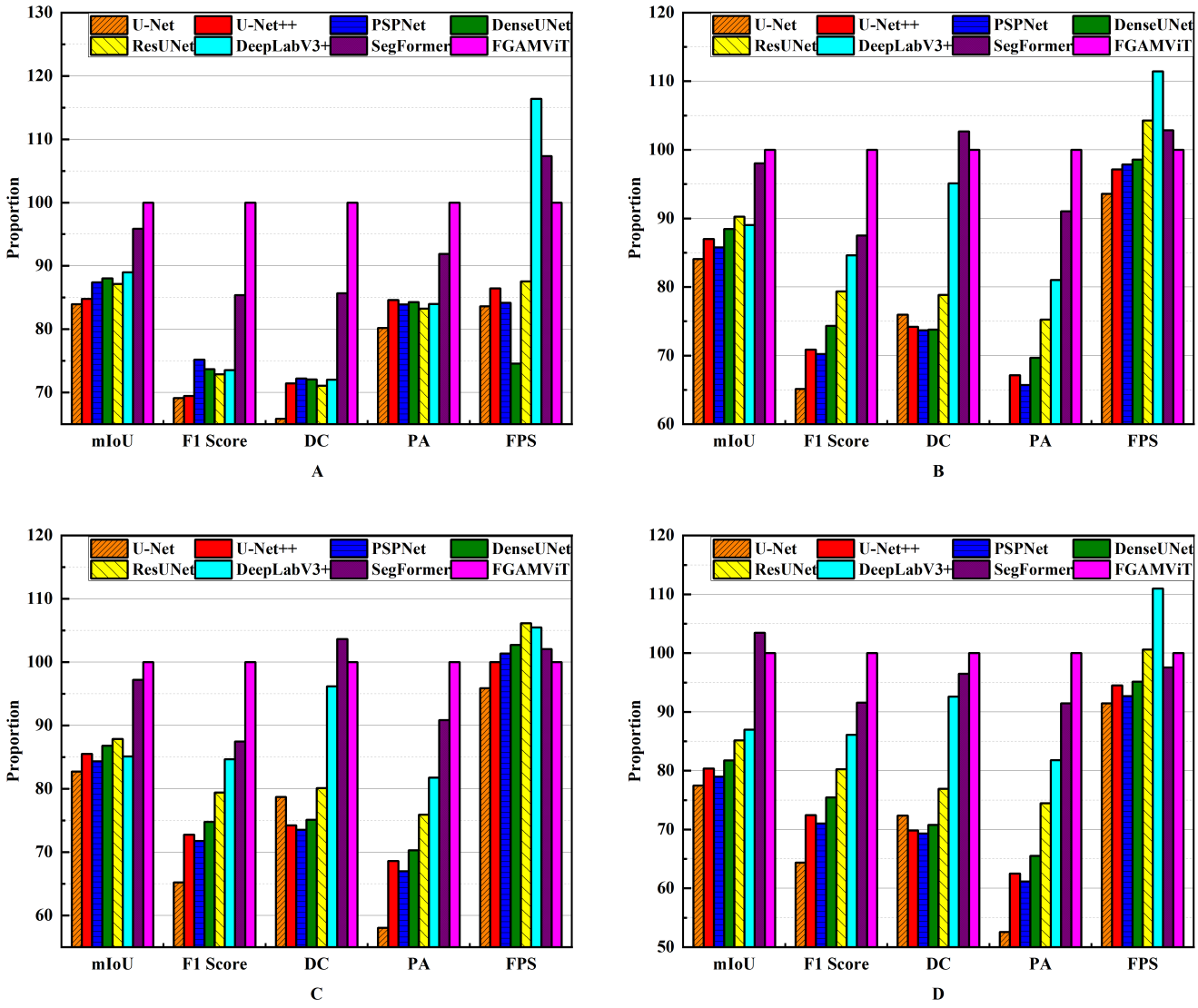

Fig. 3. Proportional Performance Comparison of Baseline Models

dataset is selected as the evaluation benchmark, and comparative results between the complete FGAMViT architecture and an ablated version (without discriminator guidance) are presented in Table II. The analysis focuses on three core metrics: mIoU, F1 Score, and Dice Coefficient.

The results reveal that the inclusion of the discriminator network significantly improves segmentation accuracy across all evaluated metrics. FGAMViT with the WGAN-GP discriminator achieves a mIoU of 38.78%, compared to 35.63% without it—an improvement of over 3 percentage points. The F1 Score and Dice Coefficient also show substantial gains, affirming the value of adversarial supervision. Several key factors contribute to these improvements:

a) Structural Refinement Through Adversarial Feedback: The discriminator network introduces an adversarial loss that evaluates the realism of the generated segmentation maps at a local patch level. This feedback forces the generator to produce spatially coherent and structurally plausible segmentation masks, effectively refining building contours and eliminating spurious noise. As a result, the model is better able to delineate boundaries in densely packed or occluded urban regions.

b) Enhanced Discrimination of Small-Scale Targets: Standard pixel-wise losses (e.g., MSE or BCE) tend to prioritise dominant regions, leading to the underrepresentation of sparse or small structures. The WGAN-GP discriminator addresses this by learning high-order spatial distributions and penalizing unrealistic segmentations of minor features. Consequently, the generator is encouraged to preserve small objects, improving the detection of isolated buildings and narrow architectural elements.

c) Improved Feature Consistency Between Prediction and Ground Truth: The discriminator evaluates both real and generated segmentation maps in conjunction with the original input image. This conditional formulation ensures that the predicted output remains contextually aligned with the input, strengthening the consistency between visual content and structural interpretation. This context-aware guidance enables the model to make more informed predictions in complex environments with mixed land uses.

d) Gradient Stability and Optimization Benefits from WGAN-GP: Unlike traditional GANs, which are prone to unstable training and mode collapse, the use of Wasserstein GAN with gradient penalty (WGAN-GP) provides smooth and stable gradients throughout the training process. This stability is crucial in semantic segmentation, where pixel-wise accuracy depends on precise, continuous feature alignment. The WGAN-GP framework enhances the training dynamics,

TABLE I
QUANTITATIVE COMPARISON OF SEGMENTATION PERFORMANCE ACROSS FOUR DATASETS. BEST RESULTS ARE IN **BOLD**; SECOND-BEST ARE UNDERLINED.

Dataset	Metric	U-Net	U-Net++	PSPNet	DenseUNet	ResUNet	DeepLabV3+	SegFormer	FGAMViT
IAIL	mIoU	34.82	35.79	35.12	36.51	37.14	36.91	38.76	**41.47**
	F1 Score	55.96	61.42	60.37	63.90	67.88	72.82	77.98	**86.72**
	Dice Coefficient	58.83	56.15	55.73	56.91	60.67	73.05	**79.12**	77.00
	Pixel Accuracy	47.83	56.72	55.98	59.24	63.75	68.88	76.48	**84.43**
	FLOPs (G)	280.2	271.0	245.3	263.4	225.3	201.6	**96.40**	98.89
	FPS	14.8	15.1	15.3	15.5	16.2	**17.5**	16.0	15.9
SpaceNet	mIoU	31.12	32.20	31.75	32.75	33.41	32.96	36.29	**37.02**
	F1 Score	49.85	54.23	53.72	56.88	60.71	64.74	66.94	**76.51**
	Dice Coefficient	51.94	50.73	50.38	50.44	53.90	65.02	**70.21**	68.37
	Pixel Accuracy	42.75	50.23	49.17	52.16	56.31	60.62	68.13	**74.83**
	FLOPs (G)	237.5	229.4	208.9	234.2	198.7	179.4	**85.80**	88.31
	FPS	13.1	13.6	13.7	13.8	14.6	**15.6**	14.4	14.0
DBE	mIoU	33.24	34.36	33.88	34.88	35.31	34.20	39.06	**40.18**
	F1 Score	53.47	59.67	58.84	61.32	65.12	69.43	71.74	**82.01**
	Dice Coefficient	56.90	53.66	53.14	54.28	57.91	69.52	**74.92**	72.28
	Pixel Accuracy	45.61	54.98	53.67	56.33	60.82	65.51	72.81	**80.13**
	FLOPs (G)	266.8	257.5	231.8	250.7	212.4	191.9	**91.80**	94.12
	FPS	14.0	14.6	14.4	14.8	**15.5**	15.3	15.4	14.9
WHUB	mIoU	30.05	31.17	30.63	31.70	33.02	33.74	**40.12**	38.78
	F1 Score	51.92	58.44	57.30	60.86	64.72	69.47	73.86	**80.67**
	Dice Coefficient	54.02	52.12	51.74	52.78	57.42	69.12	72.03	**74.64**
	Pixel Accuracy	42.15	50.13	49.02	52.54	59.72	65.57	73.33	**80.18**
	FLOPs (G)	261.9	253.1	228.7	248.2	210.8	190.5	**91.10**	93.42
	FPS	15.0	15.5	15.2	15.6	16.5	**18.2**	16.0	16.4

leading to more robust and reliable convergence.

The inclusion of a WGAN-GP-based discriminator significantly enhances the segmentation quality of FGAMViT. Through adversarial supervision, the model benefits from improved structure preservation, finer detail recovery, and enhanced boundary sharpness. These improvements are particularly valuable in high-resolution remote sensing applications, where accurate building extraction requires the integration of both local detail and global coherence. The results of this experiment demonstrate the discriminator's essential role in elevating FGAMViT's performance, making it a critical component for high-precision segmentation in complex urban environments.

TABLE II
IMPACT OF THE DISCRIMINATOR ON FGAMVIT PERFORMANCE (WHUB DATASET). THE BEST RESULTS ARE IN **BOLD**.

Configuration	mIoU (%)	F1 Score (%)	Dice(%)
w/o Discriminator	35.63	73.92	71.14
w/ Discriminator	**38.78**	**80.67**	**74.64**

D. Impact of Frequency-Domain Loss Function

Experiment 3: Evaluating the Effect of Frequency-Domain Loss on Segmentation Performance

To further investigate the contribution of individual components within FGAMViT, Experiment 3 examines the impact of incorporating a frequency-domain loss function into the generator's training objective. This experiment is conducted on the WHUB dataset, comparing two configurations: FGAMViT with and without the frequency-domain loss term L_F. The evaluation focuses on three key segmentation metrics: mIoU, F1 Score, and Dice Coefficient. Results are summarized in Table III.

The inclusion of L_F, designed to emphasise high-frequency details through 2D Fourier transform-based optimisation, leads to consistent performance improvements. FGAMViT with the frequency-domain loss achieves a mIoU of 38.78%, F1 Score of 80.67%, and Dice Coefficient of 74.64%, outperforming the model variant without L_F by 2.21%, 5.31%, and 2.79%, respectively.

Several key advantages contribute to this improvement:

a) Enhanced High-Frequency Feature Preservation: Standard spatial-domain loss functions often underweight fine-grained textures, edges, and small structures. The frequency-domain loss explicitly penalises discrepancies in high-frequency components, encouraging the network to retain sharp object boundaries and preserve fine structural cues—critical for building accurate segmentations in dense urban environments.

b) Complementarity with Adversarial and Spatial Supervision: The frequency-domain loss complements both the adversarial loss from the discriminator and the spatial MSE loss by adding a spectral-level constraint. While the adversarial loss enforces realism and the MSE maintains pixel-level accuracy, L_F bridges the gap between these two objectives by directly targeting consistency in frequency distribution. This multi-perspective supervision leads to more coherent and artifact-free predictions.

c) Better Boundary Localization and Texture Reconstruction: In urban remote sensing imagery, buildings frequently contain repetitive textures and sharp contours. The spectral sensitivity introduced by L_F allows the model to detect and reconstruct these patterns more precisely, thereby reducing over-smoothing and improving segmentation performance near complex object edges.

d) Improved Learning Stability and Convergence Speed: By providing an additional learning signal in the spectral domain, L_F helps stabilize training dynamics. The combined optimization across both spatial and frequency domains guides the network to converge toward more meaningful minima, reducing the risk of local optima associated with visually plausible but structurally inaccurate outputs.

The integration of the frequency-domain loss function into FGAMViT significantly enhances its segmentation performance, particularly in preserving detail-rich regions and maintaining boundary integrity. This enhancement is particularly beneficial in high-resolution satellite imagery, where the clarity of small objects and edges is crucial. The results confirm that L_F plays a crucial role in enhancing model precision and visual coherence, supporting its inclusion in the overall training objective.

TABLE III
IMPACT OF THE FREQUENCY-DOMAIN LOSS FUNCTION ON FGAMVIT PERFORMANCE (WHUB DATASET). THE BEST RESULTS ARE IN **BOLD**.

Configuration	mIoU (%)	F1 Score (%)	Dice (%)
w/o Frequency Loss	36.57	75.36	71.85
w/ Frequency Loss	**38.78**	**80.67**	**74.64**

VI. CONCLUSION

This study presents FGAMViT, an enhanced Transformer-based architecture tailored for high-resolution remote sensing image segmentation. Building upon the MViT framework, the proposed model introduces three critical innovations—residual pooling connections, adversarial supervision via a WGAN-GP discriminator, and a frequency-domain loss function—to overcome the limitations of conventional Transformer models in capturing fine-grained spatial details and semantic consistency in complex remote sensing scenarios.

The integration of residual pooling connections into the attention mechanism addresses the challenge of feature degradation during downsampling by preserving critical spatial information. This enhancement enables the model to maintain detailed object boundaries while ensuring computational tractability. Simultaneously, the inclusion of a discriminator network based on the WGAN-GP framework introduces adversarial feedback that significantly sharpens segmentation outputs, particularly in structurally complex or sparsely represented regions such as small buildings or narrow architectural elements.

In parallel, the incorporation of a frequency-domain loss function complements spatial supervision by guiding the generator to preserve high-frequency content. This dual-domain training strategy enhances texture reconstruction and boundary fidelity, both of which are crucial for accurate building extraction and urban feature delineation in remote sensing imagery.

Extensive experiments conducted across four benchmark datasets—IAIL, SpaceNet, DeepGlobe, and WHUB—demonstrate that FGAMViT consistently achieves superior or second-best performance across all core metrics, including mIoU, F1 Score, Dice Coefficient, and Pixel Accuracy. Despite a modest trade-off in FPS due to its enhanced feature processing, the model maintains a favourable balance between segmentation accuracy and computational efficiency.

The proposed contributions highlight the importance of integrating multiscale spatial encoding, adversarial learning, and frequency-aware optimization in designing robust segmentation models. FGAMViT establishes a new benchmark for precision in high-resolution remote sensing image segmentation, providing a flexible foundation for future work. Moving forward, these insights can be applied to further optimise Transformer-based models for real-time geospatial analysis, multimodal remote sensing fusion, and large-scale land cover mapping.

REFERENCES

[1] Q. He, X. Sun, W. Diao, Z. Yan, F. Yao, and K. Fu, "Multimodal remote sensing image segmentation with intuition-inspired hypergraph modeling," *IEEE Transactions on Image Processing*, vol. 32, pp. 1474–1487, 2023.

[2] L. Wu, M. Lu, and L. Fang, "Deep covariance alignment for domain adaptive remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[3] L. Ding, K. Zheng, D. Lin, Y. Chen, B. Liu, J. Li, and L. Bruzzone, "Mp-resnet: Multipath residual network for the semantic segmentation of high-resolution polsar images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[4] T. Liu, T. Chen, R. Niu, and A. Plaza, "Landslide detection mapping employing cnn, resnet, and densenet in the three gorges reservoir, china," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 11 417–11 428, 2021.

[5] Q. Yang, Z. Wang, S. Liu, and Z. Li, "Research on improved u-net based remote sensing image segmentation algorithm," in *2024 6th International Conference on Internet of Things, Automation and Artificial Intelligence (IoTAAI)*. IEEE, 2024, pp. 686–689.

[6] Z. Su, W. Li, Z. Ma, and R. Gao, "An improved u-net method for the semantic segmentation of remote sensing images," *Applied Intelligence*, vol. 52, no. 3, pp. 3276–3288, 2022.

[7] I. Dimitrovski, V. Spasev, S. Loshkovska, and I. Kitanovski, "U-net ensemble for enhanced semantic segmentation in remote sensing imagery," *Remote Sensing*, vol. 16, no. 12, p. 2077, 2024.

[8] Y. Xu, H. Wei, M. Lin, Y. Deng, K. Sheng, M. Zhang, F. Tang, W. Dong, F. Huang, and C. Xu, "Transformers in computational visual media: A survey," *Computational Visual Media*, vol. 8, pp. 33–62, 2022.

[9] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[10] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[11] J. Gu, H. Kwon, D. Wang, W. Ye, M. Li, Y.-H. Chen, L. Lai, V. Chandra, and D. Z. Pan, "Multi-scale high-resolution vision transformer for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 094–12 103.

[12] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "Hrformer: High-resolution vision transformer for dense predict," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7281–7293, 2021.

[13] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2998–3008.

[14] D. Yang, Z. Li, Y. Xia, and Z. Chen, "Remote sensing image super-resolution: Challenges and approaches," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*. IEEE, 2015, pp. 196–200.

[15] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding unet for remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[16] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI Brainlesion Workshop*. Springer, 2021, pp. 272–284.

[17] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proceedings of*

the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6824–6835.

[18] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4804–4814.

[19] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[20] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[21] J. Song, S. Gao, Y. Zhu, and C. Ma, "A survey of remote sensing image classification based on cnns," *Big Earth Data*, vol. 3, no. 3, pp. 232–254, 2019.

[22] M. Alam, J.-F. Wang, C. Guangpei, L. Yunrong, and Y. Chen, "Convolutional neural network for the semantic segmentation of remote sensing images," *Mobile Networks and Applications*, vol. 26, pp. 200–215, 2021.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[25] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," *Applied Sciences*, vol. 12, no. 18, p. 8972, 2022.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances In Neural Information Processing Systems*, vol. 30, 2017.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv Preprint ArXiv:2010.11929*, 2020.

[28] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are rnns: Fast autoregressive transformers with linear attention," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5156–5165.

[29] S. Jaszczur, A. Chowdhery, A. Mohiuddin, L. Kaiser, W. Gajewski, H. Michalewski, and J. Kanerva, "Sparse is enough in scaling transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9895–9907, 2021.

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[31] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.

[32] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 214–223.

[33] U. Demir and G. Unal, "Patch-based image inpainting with generative adversarial networks," *ArXiv Preprint ArXiv:1803.07422*, 2018.

[34] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017.

[35] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, "Spacenet: A remote sensing dataset and challenge series," *ArXiv Preprint ArXiv:1807.01232*, 2018.

[36] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 172–181.

[37] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 2018.

[38] Y. Cao, S. Liu, Y. Peng, and J. Li, "Denseunet: densely connected unet for electron microscopy image segmentation," *IET Image Processing*, vol. 14, no. 12, pp. 2682–2689, 2020.

[39] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," in *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2019, pp. 225–2255.

[40] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.