IEEE Open Journal of the Computer Society

Received 7 March 2025; revised 23 April 2025; accepted 25 April 2025. Date of publication 28 April 2025; date of current version 16 May 2025. The review of this article was arranged by Associate Editor Peng Li.

Digital Object Identifier 10.1109/OJCS.2025.3565185

SwinSegFormer: Advancing Aerial Image Semantic Segmentation for Flood Detection

MUHAMMAD TARIQ SHAHEEN [1], HAFSA IQBAL [2], NUMAN KHURSHID [1], HALEEMA SADIA [3], AND NASIR SAEED [3] (Senior Member, IEEE)

[1]School of Electrical Engineering and Computer Science, National University of Scienes and Technology (NUST), 44000 Islamabad, Pakistan
[2]Autonomous Mobility and Perception Laboratory (AMPL), University Carlos III of Madrid, 28911 Madrid, Spain
[3]Department of Electrical and Communication Engineering, United Arab Emirates University, Al Ain 15551, UAE

CORRESPONDING AUTHORS: HAFSA IQBAL; NASIR SAEED (e-mail: hiqbal@ing.uc3m.es; mr.nasir.saeed@ieee.org).

This work was supported by the United Arab Emirates University (UAEU) through the UAEU Program for Advanced Research (UPAR) under Grant 12N174.

ABSTRACT Semantic segmentation of aerial images is essential for unmanned aerial vehicle (UAV) applications in disaster management, particularly for identifying the flood-affected areas. Traditional techniques face challenges in capturing global semantic information due to their limited receptive fields, and high computational requirement. To address these issues, we propose a novel transformer-based model named SwinSegFormer, which feature a hierarchical encoder that efficiently generates multi-scale high-resolution features along with a lightweight decoder to reduce computational overhead. The proposed model is trained on FloodNet dataset and demonstrates efficient performance on challenging classes such as vehicles, pools, and flooded and non-flooded roads, which are crucial for effective disaster management. Additionally, we developed a post-processing module to categorize areas into flooded and non-flooded. The model achieves a validation mIoU of 75.1%, mDice of 85.4%, and mACC of 87.1%, representing a 10-12% improvement over state-of-the-art vision transformer-based methods. The effectiveness of model is further evaluated on real-world unlabeled flood imagery, highlighting its potential for supporting first aid activities during floods. Relevant codes are available at: https://github.com/Shaheen1998/SwinSegFormer.

INDEX TERMS Flood detection, semantic segmentation, SegFormer, swin transformer, unmanned aerial vehicles (UAVs), vision transformers.

I. INTRODUCTION

Semantic segmentation deployed on Unmanned Aerial Vehicles (UAVs) plays a vital role in natural disaster management, particularly in formulating emergency rescue strategies through real-time analysis of affected areas. Traditional disaster assessment methods, relying on manual surveys and satellite imagery, are resource-intensive and time-consuming, lacking the precision necessary for effective decision-making. In contrast, UAVs equipped with onboard computing devices enable real-time semantic segmentation, allowing rapid identification of flood-affected areas and damage assessment. This capability is especially critical in flash flood scenarios, where expeditious response is crucial to minimize damage (as shown in Fig. 1). These capabilities are especially critical in flash floods, where rapid response is essential to minimize damage.

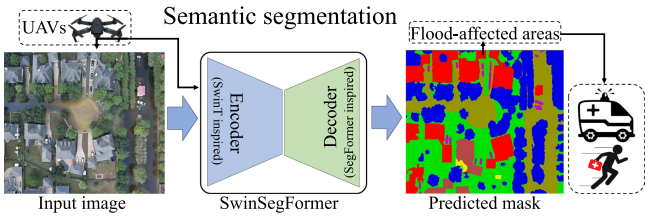

FIGURE 1. The applications of semantic segmentation to aerial imagery are truly remarkable for disaster management. It allows identifying the flood-affected areas using the computing devices installed on UAVs and conduct life-saving flood relief activities, such as distributing first-aid kits and rescuing survivors.

UAVs capture Very High-Resolution (VHR) aerial imagery [1], [2], providing detailed information for precise classification, and localization of affected areas, particularly

in urban environment. This technology improves situational awareness by accurately segmenting flood-impacted areas, facilitates efficient resource allocation, and optimizes rescue operations. The combination of semantic segmentation and UAV technology improves the disaster management by overcoming the limitations of traditional assessment methods, offering a scalable and adaptive solution for real-world emergency situations.

Various CNN-based models for semantic segmentation of satellite imagery have been proposed, such as HRNet [3], VGGNet [4], ResNet [5], and DeepLabV3+ [6]. These models incorporate attention mechanisms [7], [8], [9], [10], [11], combine global and local information and vary network depth and width. However, traditional CNN-based models still have limitations in remote sensing imaging (RSI), such as difficulties in capturing long-range dependencies and recovering lost spatial information in deeper layers due to their local receptive fields. To address these limitations, Vision Transformers (ViTs) [12] introduced a ViT-based backbone for semantic segmentation. Although SETR [13] achieves superior performance, it requires high computational costs and produce low-resolution feature representations.To address these challenges, SwinT [19] introduces a hierarchical architecture that efficiently models long-term dependencies and produces multi-scale resolution features. However, its decoder head comprises multiple transformer layers, making it complex and computationally expensive. The SegFormer [12] model reduces computational cost but encounter challenges while accurately segmenting certain narrow classes. Therefore, this research primarily addresses the limitations of state-of-the-art (SOTA) models in semantic segmentation for satellite and UAV-based imagery, particularly focusing on mitigating the effects of natural disasters like floods. Leveraging the advantages of Swin Transformer (SwinT) and the SegFormer, we propose an hybrid model, i.e., *SwinSegFormer*. The significant contributions of this work are as follows:

- We propose a novel SwinSegFormer model that leverages the strengths of the SwinT hierarchical encoder, which efficiently models long-term dependencies and multi-resolution features, with a SegFormer decoder to maintain computational efficiency while enhancing segmentation precision.
- We apply SwinSegFormer to flood detection applications employing the FloodNet dataset. This detection facilitates the estimation of flood-affected areas using semantic segmentation and classifying regions into flooded and non-flooded, enabling more effective planning and execution of first-aid activities.
- We evaluated SOTA CNNs and transformer models, i.e., DeepLabV3+, SwinT, various SegFormer variants, and their hybrid combinations to assess the effectiveness in remote sensing imagery (RSI) segmentation tasks.
- To assess the real-world applicability of the proposed model, we make inference on low-resolution, real-world unlabled flood images from regions such as California, Mississippi, Brick Township, Hurricane Sally in the USA, and floods in Pakistan. These images were classified into flooded and non-flooded regions, showcasing the model's effectiveness in real-world scenarios.

The rest of the article is organized as follows: Section II discusses CNN and transformer based semantic segmentation techniques applied in RSIs and review their suitability for flood detection. Section III details SwinT backbone, lightweight decoder head of the SegFormer model, and the hybrid architecture of SwinSegFormer model. Section IV discusses the dataset used in this research, implementation details for training, comprehensive ablation studies, and comparative analysis of the SwinSegFormer model with SOTA models to validate our proposed model. Finally, Section V concludes our work and highlights future directions.

II. LITERATURE REVIEW

The following subsections delve into the literature for semantic segmentation of RSIs. We explore the CNN-based semantic segmentation techniques and transformer-based semantic segmentation techniques applied in RSIs. Theoretical comparison of several SOTA models is presented in Table 1.

A. SEMANTIC SEGMENTATION IN THE FIELD OF RSIS

RSIs are significant in many applications, including land cover mapping, crop yield estimation [13], and natural disaster management systems [20]. The accuracy of semantic segmentation in RSIs is vital for these applications. Deep learning techniques have demonstrated promising results in semantic segmentation of remote imagery, with approaches such as [21], [22] being introduced to tackle both inter-class and intra-class variations in remote images.

Recognizing the significance of extracting contextual information from images is crucial to optimizing the semantic segmentation in RSIs. In this regard, Convolutional Networks [23] and Fully Convolutional Networks (FCNs) [4] have enhanced the receptive field by incorporating pooling layers to capture contextual image information. However, the multiple down-sampling steps in pooling layers lead to the loss of fine-grained details. FCNs incorporate skip connections to mitigate this issue to establish data flow between shallow and deep features. Additionally, an encoder-decoder architecture is employed to restore feature map resolution, effectively recovering information lost during the pooling process. In [24], ResNet was introduced as a residual structure to tackle the vanishing gradient problem. The authors proposed ResNeXt [25], which replaces ResNet's residual blocks with grouped convolutions and expands the network to reduce the number of parameters. Among these architectural advancements, the focus was on harnessing low-level resolution features to extract high-resolution semantic features. The study introduces HRNet [26], specifically designed to preserve higher-resolution features and generate parallel feature maps for semantic segmentation from the highest to lowest resolutions at all stages. This facilitates multiple information exchanges for recurrent multi-resolution features. The primary role of the decoder is to restore feature resolution and fully exploit multi-scale feature mapping to produce

TABLE 1. Theoretical Comparison of the Proposed Model With SOTA Models, Highlighting Their Distinctive Attributes

Features/ Models	Backbone	Encoder/ Decoder	Attention mechanism	Positional encoding	Feature resolutions	Receptive field
CNNs [5]	Convolution	Simple/Simple	No	Free	Low	Limited
ViTs [14]	Pure transformer	Non-hierarchical/Complex	Self-attention	Yes	Single-scale	Global
SETR [15]	Pure transformer	Non-hierarchical/Complex	Self-attention	Yes	Single-scale	Global
PVT [16]	PVT	Hierarchical/Heavy	SRA	Yes	Multi-scale	Global
CapMatch [17]	Transformer capsule	Supervised, and unsupervised	Capsule attention	Yes	Multi-scale	Global
DKN [18]	ResMulti-Trans	Densely-Connected/Heavy	DDSD	Yes	Multi-scale	Global
SwinT [19]	SwinT	Hierarchical/Heavy	W-MSA, SW-MSA	Yes	Multi-scale	Global
SegFormer [12]	PVT	Hierarchical/Lightweight	ESA	Free	Multi-scale	Global
SwinSegFormer	SwinT	Hierarchical/Lightweight	W-MSA, SW-MSA	Yes	Multi-scale	Global

semantic segmentation results. Commonly used decoders include architectures like symmetrical auto-decoders [27] and Deeplabv3+ [28]. Notable early robust encoder-decoder structures includes DeconvNet [29], SegNet [30], U-Net [31], and others. Among these, U-Net variants have gained prominence due to their effective fusion of multi-scale resolution feature maps. The U-shaped decoder [32], [33], [34], in particular, excels in retaining low-level resolution features to generate essential local data. Innovative U-shaped decoder variants, such as res-unet [32] and dense-unet [33], draw inspiration from ResNet and DenseNet [35], respectively, to improve the performance corresponding to each module of the U-Net architecture.

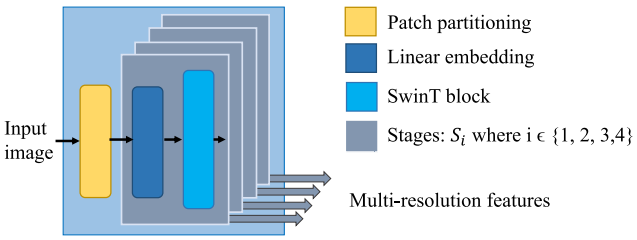

FIGURE 2. SwinT architecture [19] comprises a hierarchical encoder that operates across four stages S_i which generate multi-resolution feature maps y_i. Each stage comprises two main components, linear embedding block, and patch merging SwinT block.

B. TRANSFORMER-BASED SEMANTIC SEGMENTATION TECHNIQUES

The transformer-based network was initially invented for natural language processing and quickly garnered the research community's attention due to its versatile architecture [36]. A transformer network is a deep learning (DL) sequential architecture that employs a self-attention mechanism to extract the intrinsic features. ViTs [14] proposed a pure transformer backbone capable of replacing CNNs and acquiring SOTA performance in various vision-related tasks. This research was revolutionary as it utilized a pure transformer architecture for tasks like image classification, segmentation, and recognition. While ViTs achieved promising results, they faced limitations in low-resolution feature mapping and exponentially increasing complexity with larger image sizes, making them unsuitable as a general backbone for dense vision tasks. The SETR [15] adopted the ViTs backbone for sequence-to-sequence prediction tasks. It employs a pure transformer as an encoder and integrates a simple decoder based on CNN architecture, achieving impressive performance compared to all existing SOTA models. However, ViTs suffered two significant disadvantages: they generated single-scale low-resolution features and incurred high processing costs when handling large images.

The Pyramid Vision Transformer (PVT) [16] is introduced as the first transformer-based approach with a pyramid structure and also used as the backbone for semantic segmentation tasks [37]. PVT employs an encoder with a gradually shrinking pyramid architecture to produce high-resolution features and incorporates a spatial-reduction attention mechanism to minimize computational costs. Although they succeeded in reducing the computational costs of the encoder, they overlooked the resource-intensive decoder, which rendered the PVT computationally inefficient for processing high-resolution images.

HIPT [38] proposes a hierarchical ViT architecture to decompose gigapixel images into nested sequences of regions, patches, and visual tokens to extract bottleneck features. Fd-vit [39] introduces an improved version of hierarchical ViT that overcomes the challenge of high computational cost by employing flexible downsampling layers. Xue et al. [40] presents a deep ViT model specifically tailored for the classification of high-dimensional LiDAR data. SwinT [19] introduced a hierarchical encoder to address the challenge of single-scale low-resolution features and integrate a window-based self-attention mechanism to minimize computational costs. Nevertheless, capturing global pixel relationships requires more than relying on a single local window for self-attention. Therefore, the approach incorporates two consecutive transformer architectures with shifting windows to facilitate global self-attention computations. An overview of the SwinT architecture is described in Fig. 2. In the SwinT architecture, the process involves four distinct stages to acquire the hierarchical feature maps. While substantial enhancements were made to the encoder, the decoder was somewhat

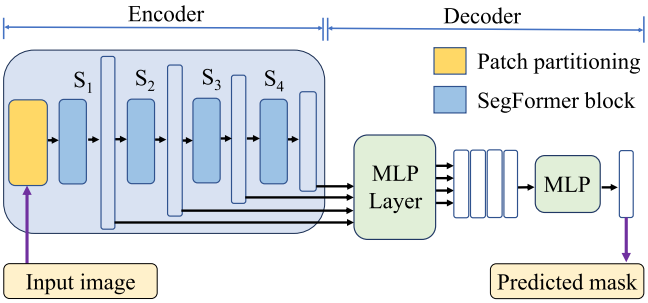

FIGURE 3. The SegFormer [12] architecture comprises of two modules: a transformer-based hierarchical encoder for multi-resolution feature extraction and a lightweight All-MLP decoder head.

overlooked, rendering it computationally impractical for processing large images.

The SegFormer model [12] introduced an innovative hierarchical encoder that produces multi-level features and a lightweight MLP decoder, effectively addressing the challenge of computationally expensive decoders. Its encoder is free from positional encoding, eliminating the requirement for positional code interpolation when dealing with images of varying resolutions, thus enhancing performance in segmentation tasks. An overview of the SegFormer architecture is depicted in Fig. 3. Encoder generates multi-level feature resolutions and passes them to the MLP decoder head, which predicts the final segmentation map. Despite addressing the challenge of computational costs, SegFormer encounters difficulties in segmenting classes with narrow masks.

The application of transformers in semantic segmentation is still in its initial evaluation phase, requiring computationally efficient architectures capable of accurately extracting high-resolution features. A transformer-based model, i.e., CapMatch [17], utilizes a contrastive transformer capsule approach with feature-based knowledge distillation to improve semi-supervised learning in human activity recognition. Similarly, the Densely Knowledge-Aware Network (DKN) [18] employs a densely connected transformer-based framework for multivariate time series classification. These models incorporate techniques like Densely Dual Self-Distillation (DDSD), which promotes mutual knowledge transfer between lower- and higher-level semantic information, helping to regularize the model and improve its representation learning performance.

Although these approaches utilize transformers to learn long-range dependencies, their methodologies differ significantly from semantic segmentation tasks. Unlike these models, our model is specifically designed for remote sensing imagery analysis. It incorporate a hierarchical transformer encoder and a lightweight decoder, creating a dual-branch network architecture. This design strikes a balance between high-resolution semantic segmentation capabilities for aerial images and computational efficiency.

C. FLOOD DETECTION USING ARIEL IMAGERY

Several strategies for flood-affected area detection aim to enhance rescue operations and emergency responses, thereby mitigating the impacts of flooding. Natural disasters like floods are very unpredictable and can cause significant damage to society and infrastructure. Particularly, flash floods allow minimal time to react, and real-time assessment is crucial for an effective disaster response and minimizing its impact. In Spain and Pakistan, the recent flood events in 2024, 2022 and 2023, highlighted the immediate need for effective techniques to detect flooded areas and improve flood disaster management. Remote sensing technologies have gained popularity due to their cost-effectiveness during flood crises, enabling early damage assessment and facilitating rescue operations. This section examines the application of remote sensing imagery (RSI) for flood detection.

In a recent study [41], researchers evaluated UNet-MobileNetV3 [41], DeepLabV3+ [28], and PSPNet [28] trained with FloodNet dataset [42]. They specifically compared the performance of these models in distinguishing flooded and non-flooded areas. The study focused on assessing the effectiveness of real-time semantic segmentation models compared to their offline counterparts, especially in challenging aerial imagery and hostile environments.

In another study [43], models such as UNetFormer [44], SegFormerB0 [12], and others were evaluated using the FloodNet dataset [42] for aerial image segmentation during crises. The study addresses the urgent requirement for real-time semantic segmentation, particularly highlighting applications involving UAVs. Among the models tested, the transformer-based SegFormerB0 [12] demonstrated superior performance. However, the study did not analyze all of the variants of SegFormer, indicating a need for further exploration. This work provides an intensive comparison of various transformer-based models, including SwinT and SegFormer, for the detection of flood-affected areas with minimizing computational resources.

III. PROPOSED METHODOLOGY

We proposed a novel transformer-based model for efficient semantic segmentation of aerial imagery, leveraging the hierarchical encoder from SwinT and the MLP decoder from SegFormer. This approach improves segmentation performance while reducing computational demands. Section III-A provides overview of SwinSegFormer architecture, SwinT and SegFormer discussed in Section III-B and III-C, respectively and Section III-D highlights the estimation of flood affected area.

A. SWINSEGFORMER: ARCHITECTURAL OVERVIEW

An overview of the SwinSegFormer architecture, depicted in Fig. 4, integrates SwinT's hierarchical encoder with SegFormer's lightweight decoder. The proposed model incorporates the following key components:

1) Instead of using a traditional CNN backbone, we opted for the SwinT backbone as our feature extraction method due to its hierarchical architecture, which

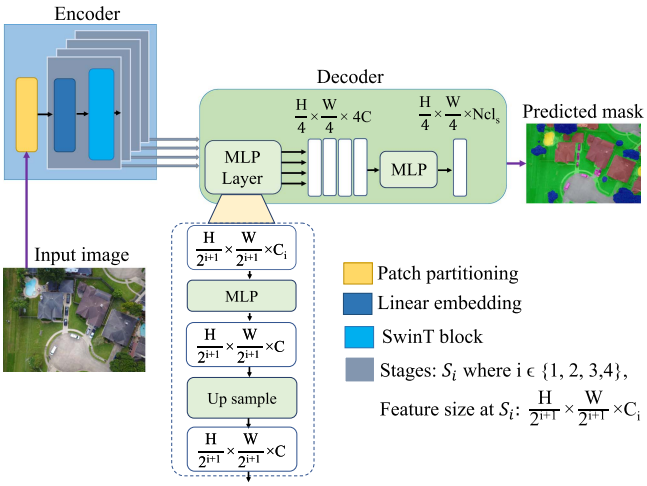

FIGURE 4. SwinSegFormer Architecture: Encoder comprises of four stages (S_i), and generates multi-resolution features y_i, and used to predict the segmentation mask from decoder.

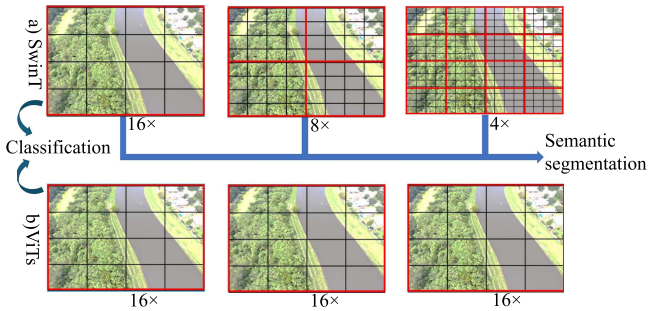

FIGURE 5. Multi-resolution features extraction: (a) SwinT [19] builds a hierarchical feature map with different resolutions by merging patches (black lines) in deeper layers, and compute self-attention within local windows (red lines) and (b) ViTs [14] extract only fixed-resolution features.

accurately models long-term image dependencies without introducing inductive bias. The SwinT provides multi-resolution features at different stages and reduces computational resources by using a window-based self-attention mechanism limited to local windows.

2) We used the lightweight decoder from the SegFormer model to reconstruct the feature map. This decoder consists only of MLP layers, avoiding computationally complex components. We chose this because our hierarchical encoder already extracts multi-resolution features and employs a shifting window-based self-attention mechanism, providing a more extensive receptive field than traditional encoders.
3) Our proposed SwinSegFormer model integrates a dual network design: leveraging the SwinT hierarchical encoder for multi-resolution feature generation and an MLP decoder head for progressive feature refinement and reducing computational overhead, enhancing semantic segmentation performance in aerial imagery. Additionally, we have developed a post-processing module to calculate flood-affected areas and categorize them into flooded and non-flooded regions.

The encoder initially processes an RGB input image of dimensions $H \times W \times C$, where H, W, C represents height, width, and channels ($C = 3$). It splits the input image into patches and treats each patch as a token, utilizing a patch-splitting module. The SwinT-based hierarchical encoder operates across four stages S_i to generate feature maps y_i with different resolutions, i.e., $\{y_1, y_2, y_3, y_4\}$. Each stage comprises two key components: patch merging and SwinT. Patch merging divides the image into non-overlapping patches using a moving window and projects them through linear embedding. Stage 1 (when $i = 1$) produces output feature maps of size $H/4 \times W/4$. Subsequent stages provides high resolution feature maps with following sizes; $H/8 \times W/8$ (stage 2), $H/16 \times W/16$ (stage 3), and $H/32 \times W/32$ (stage 4). Finally, the MLP decoder combines and refines feature resolutions from all stages to generate and predict the segmentation mask. The following subsections provide a mathematical formulation of the SwinSegFormer encoder-decoder architecture.

B. SWINT BACKBONE

The SwinT architecture, featuring the SwinT block as its backbone, efficiently captures global spatial dependencies in images and produces multi-resolution features crucial for tasks such as semantic segmentation. Unlike traditional transformers in ViTs that use Multi-head Self-Attention (MSA) globally, resulting in quadratic computational complexity, SwinT introduces Window-Based Multi-Head Self-Attention (W-MSA) and Shifting-Window-Based Multi-Head Self-Attention (SW-MSA) modules. These innovations replace standard MSA, effectively enhancing the model's ability to capture global contextual information and spatial dependencies within images. For a detailed discussion on MSA in standard transformer architectures, refer to [14], [45]. Following (1) and (2) explains how SwinT scales the quadratic complexity of MSA (Ω_{MSA}) into linear complexity using W-MSA (Ω_{W-MSA});

$$\Omega_{MSA} = 4hwC^2 + 2(hw)^2C, \tag{1}$$

where hw is the patch size, which increases quadratically with an image of size ($H \times W$) and C dimension. Therefore, the (1) is modified by introducing a fixed window size (M) which inturn scales linearly the computational cost of the module (Ω_{W-MSA}) as follows;

$$\Omega_{W-MSA} - \Omega_{MSA} = 4hwC^2 + 2M^2hwC, \tag{2}$$

where M is the window size (default sets to $M = 7$) and each window comprises of the patch (hw), as shown in Fig. 5. This linear scaling enables SwinT to process high-resolution images with relatively low computational costs efficiently. Self-attention calculations are confined to local windows only to further reduce the computational cost. Here, a shifted window partitioning strategy is proposed that switches between two window partitioning configurations in subsequent SwinT blocks introduce cross-window connections while preserving the effective computation of nonoverlapping windows. When

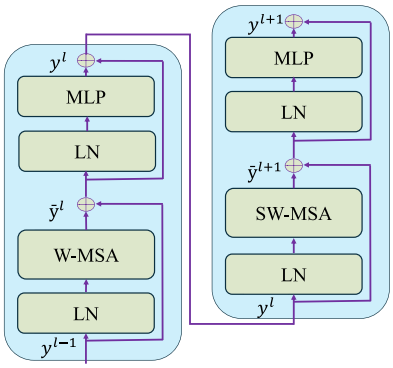

FIGURE 6. The key modules of SwinT [19] are W-MSA and SW-MSA. Features are extracted from W-MSA and then processed through SW-MSA block. Here, LN is Normalization Layer, and $\bar{y}^l$ represents the output characteristics of W-MSA module for layer l.

using shifted window partitioning, it is important to note that this approach can generate more windows, some of which may be smaller than M in size. However, one simple solution to overcome this limitation is introducing padding in smaller windows and masking the padded values during attention computation.

Fig. 6 illustrates the key modules of SwinT, specifically W-MSA and SW-MSA, where the features are extracted from W-MSA and subsequently processed through the SW-MSA block. Successive W-MSA and SW-MSA blocks enable SwinT to effectively capture the global spatial dependencies in images via cross-window connections. Mathematically, the feature map generation can be described as follows;

$$\bar{y}^l = W_{MSA}\left(LN\left(y^{(l-1)}\right)\right) + y^{(l-1)}, \tag{3}$$

where $\bar{y}^l$ represents the output characteristics of the W-MSA module in layer l. This equation entails applying W-MSA to the feature map y^{l-1} following the Layer Normalization (LN) to normalize the results. Normalized results are then added to the feature map y^{l-1} obtained in the previous layer $l-1$, establishing a residual connection among them. $\bar{y}^l$ is then passes through MLP to obtain y^l as follows;

$$y^l = MLP\left(LN\left(\bar{y}^l\right)\right) + \bar{y}^l. \tag{4}$$

As (4) depicts, this computation involves applying LN to $\bar{y}^l$ and processing it with an MLP. The result is then added to the $\bar{y}^l$ and establishes a residual connection. y^l is then employed to calculate the $\bar{y}^{(l+1)}$ by processing it through SW-MSA module for layer $(l+1)$, as follows;

$$\bar{y}^{(l+1)} = \text{SW-MSA}\left(LN\left(y^l\right)\right) + y^l. \tag{5}$$

Similar to (4), $\bar{y}^{(l+1)}$ is then passes through MLP as;

$$y^{(l+1)} = MLP\left(LN\left(\bar{y}^{(l+1)}\right)\right) + \bar{y}^{(l+1)}. \tag{6}$$

These equations highlight the intricate feature map transformations and interactions within the SwinT blocks, incorporating self-attention, normalization, and residual connections for effective feature learning. This allows the SwinT to merge the patches and provides a hierarchical feature map y_i with a resolution of $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$ corresponding to each input image, where i indicates the resolution levels $\{1, 2, 3, 4\}$. The extraction of multi-resolution features improves semantic segmentation performance by providing fine-grained low-resolution features and high-resolution coarse features.

C. MLP DECODER

A simplified decoder inspired by the SegFormer model [12] is proposed, emphasizing reducing the computational complexity while improving the architecture's performance. Our hierarchical encoder integrates shift and cross-window-based self-attention mechanisms, providing an extensive receptive field, making this simple decoder well-suitable. This decoder employs an all-MLP architecture for decoding purposes, which comprises four essential steps to reconstruct feature maps and generate semantic segmentation representations of aerial images. These steps can be mathematically formulated as follows:

1. Unifying channel dimensions: involves unifying the channel dimensions across different feature levels generated by the encoder. Each feature map y_i is passed through MLP;

$$y_i' = \text{MLP}(C_i, C)(y_i), \quad \forall i, \tag{7}$$

where C_i and C represent the dimensions of the input and output feature vector, respectively.

2. Upsampling: The obtained features y_i' are upsampled to $1/4^{th}$ of their original size as follows;

$$y_i'' = \text{Upsample}\left(\frac{W}{4} \times \frac{W}{4}\right)(y_i'), \quad \forall i. \tag{8}$$

3. Channel fusion: The upsampled features y_i'' are concatenated and processed through an additional MLP;

$$y = \text{MLP}(4C, C)(\text{Concat}(y_i')), \quad \forall i. \tag{9}$$

This step combines information from multiple feature levels, thereby enhancing the model's ability to capture contextual information.

4. Segmentation mask prediction: Segmentation mask M is predicted by applying MLP from fused features y;

$$M = \text{MLP}(N_{\text{cls}}, C)(y), \tag{10}$$

where M is segmentation mask with a resolution of $W/4 \times W/4 \times N_{\text{cls}}$, where N_{cls} represents the number of categories. The decoding process is concluded with this step and generates the segmentation mask prediction.

D. ESTIMATION OF FLOOD-AFFECTED AREA

We have developed a post-processing module to analyze predicted masks, categorizing regions into flooded and non-flooded areas. Initially, the module identifies all predicted classes within the predicted mask and then determines the

TABLE 2. Distribution of Classes in FloodNet dataset [42]

Classes	Images (%)	Instances (%)
Building(F)	245 (10.47%)	3248 (9.01%)
Building(NF)	880 (37.59%)	3427 (9.50%)
Road(F)	264 (11.29%)	495 (1.37%)
Road(NF)	1175 (50.20%)	2155 (5.98%)
Vehicle	813 (34.70%)	4535 (12.58%)
Pool	531 (22.67%)	1141 (3.16%)
Tree	1885 (80.48%)	19682 (54.60%)
Water	984 (42.02%)	1374 (3.81%)

presence of flooded classes, such as flooded roads or flooded buildings. If any flooded class exists, the module estimate the area covered by that flooded class, if this exceeds 30% of the image area, the predicted mask is labeled as flooded; otherwise, it is classified as non-flooded.

IV. EXPERIMENTAL SETUP AND RESULTS

In this section, we present the results of our experiments and analyses, to compare the performance of the SwinSegFormer model with the SOTA models designed for semantic segmentation in remote-sensing images. Our evaluation primarily focuses on the FloodNet dataset [42], for which we provide an introduction about the dataset, discuss the evaluation metrics and present implementation details. The results include quantitative and qualitative analyses, and ablation studies to assess the model's effectiveness for semantic segmentation of remote sensing images.

A. DATASET

1) FLOODNET DATASET WITH LABELLED AND UNLABELED IMAGES

FloodNet [42] is a unique pixel-level training dataset of 6777 high-resolution (3000×4000) labelled and unlabeled images, including 2343 pixel-level annotated images. These images were captured by DJI Mavic Pro quadcopters after Hurricane Harvey in 2017, providing a spatial resolution of 1.5cm from an altitude of 200 feet. This resolution distinguishes FloodNet datasets from other datasets having lower resolutions. Unlike many flood detection datasets rely on satellite imagery with limited spatial resolution and infrequent revisits, FloodNet leverages UAVs to capture high-resolution images swiftly in disaster-affected regions, making it suitable for computer vision applications. Image annotations allows to identify the flooded and non-flooded areas, additionally classifying the flooded area into flooded-buildings and flooded-roads, aids in assessing the extent of damage. Moreover, a 'water' class is introduced to differentiate the natural water bodies from flooded areas, and images classified as 'flooded' if floodwater covers over 30% of the image. However, accurate segmentation of such images is challenging due to imbalances in training image distribution and instances (depicted in Table 2), particularly for challenging classes like vehicles, pools, and flooded-roads.FloodNet captures detailed aerial data, highlighting the challenges associated with semantic segmentation in UAV applications. Our analysis of FloodNet focuses on image segmentation, which is crucial for identifying disaster-affected regions. In the aftermath of a disaster, the response teams rely on classifying impacted areas (e.g., floods) and conduct semantic segmentation to assess the situation and estimate the damage. Our research aligns with these requirements and aims to collaborate with response and rescue teams.

2) MODEL VALIDATION WITH UNLABEL DATASET

We have compiled a small unlabeled dataset of real-world images comprising flooded and non-flooded areas. The dataset covers recent flood events, including the 2022 and 2024 floods in Pakistan, coastal storm in Brick Township, New Jersey, in 2024, record-breaking rainfall in Charleston, South Carolina, USA, in 2024, slow movement of Hurricane Sally leading to flash floods in United States (AL, FL, MS) in 2020, and a major flood along the Mississippi River in 2019, which affected multiple states including Louisiana, USA. These images were employed to validate the performance of our model and collected from the Internet. Despite having lower resolutions compared to training dataset, i.e., FloodNet, they remain significant for real-world applications and serve their purpose in assessing the robustness of models, especially in detecting flooded areas.

B. EVALUATION METRICS

We evaluated the models' performance using various evaluation metrics, which include Intersection over Union (IoU), Dice Score, F1-score, Precision, Recall, Overall Accuracy (OA), and the means for all these metrics. The following equations formally describe these evaluation metrics:

IoU (11) measures the overlap between predicted (N_p) and ground truth (N_{gt}) regions as follows;

$$IoU = \frac{N_p \cap N_{gt}}{N_p \cup N_{gt}}. \tag{11}$$

F1-score in (12) that is the harmonic mean of the precision and Recall, computed as follows;

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{12}$$

Dice score in (13) measures spatial overlap between predicted and ground truth regions and is computed as;

$$\text{Dice} = \frac{2 \times TP}{(TP + FP) \cdot (TP + FN)}. \tag{13}$$

C. EXPERIMENTAL SETTINGS

We evaluated the performance and efficiency of various semantic segmentation methods on FloodNet (as shown in Table 5). We used a data augmentation pipeline to broaden the training and validation datasets. The training pipeline included random resizing and cropping to 512×512 pixels, horizontal flipping with a probability of 0.5, and photometric distortions for brightness, saturation, contrast, and hue. Our

TABLE 3. Performance Evaluation of Proposed Model With SegFormer, SwinT and Other Hybrid Variants (encoder+decoer)

	SwinSegFormer (Proposed)		SwinT [19]		SegFormer [12]		SwinT+ViT		SegFormer+ViT		SegFormer+SwinT	
Class	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
Building(F)	77.7	87.4	71.8	83.6	68.6	81.4	78.95	87.23	69.53	82.03	69.49	82.0
Building(NF)	78.6	88.1	77.7	87.4	62.8	77.2	80.43	89.15	56.71	72.37	63.4	77.6
Road(F)	63.5	77.6	58.7	74.0	50.3	66.9	60.71	75.55	37.91	54.98	53.35	69.58
Road (NF)	82.5	90.4	78.3	87.9	71.7	83.5	83.15	90.10	70.64	82.79	73.09	84.45
Water	72.0	83.7	67.6	80.7	60.4	75.3	70.26	82.53	62.21	76.71	64.85	78.68
Tree	85.1	91.9	73.6	84.8	75.5	86.0	84.84	91.80	76.5	86.69	78.09	87.7
Vehicle	55.2	68.6	30.9	47.2	18.6	31.3	52.55	68.89	12.56	22.31	19.72	32.94
Pool	65.5	79.2	49.6	66.3	37.5	54.5	64.40	78.35	40.03	57.17	35.74	52.66
Grass	89.1	94.2	85.3	92.1	83.0	90.7	88.42	93.85	83.86	91.22	83.33	90.91
Mean	75.2	85.4	64.1	75.2	60.4	70.7	73.08	84.62	55.89	66.49	60.93	77.83

experiments were implemented with Python Version 3.7.12, CUDA Version 11.4, and PyTorch Version 1.12.0. The training involved a range of 10k-to-50 k iterations, a training batch size of 2, and a validation batch size of 1. Due to the complexity of transformer-based models and the large FloodNet dataset, smaller batch sizes were chosen to manage computational resources. We used the Adam optimizer with an initial learning rate of 6e-05, betas of $(0.9, 0.999)$, and a weight decay of 0.01 and 0.01 for the SGD optimizer during training. The pre-trained weights initialized all model parameters. The FloodNet data set was divided into 60% for training, 30% for validation, and 10% for testing. GPU hardware was used to accelerate the training process. SwinSegFormer performed well with FloodNet and setting a benchmark for semantic segmentation and flood detection. We used an unlabeled collected dataset and an unlabeled FloodNet dataset to infer the SwinSegFormer model, and it effectively located flood-affected areas.

D. ABLATION STUDIES

This section evaluates the performance of our model and compares with single- and dual-branch SOTA networks, i.e., SwinT, SegFormer, hybrid combinations, and variants of SegFormer, for flood detection application [42].

1) COMPARISON OF ENCODER AND DECODER

We performed a comprehensive experimental evaluation of various model variants, combining different encoder and decoder architectures from SOTA. The study included standalone architectures (DeepLabv3+ [41], SegFormer [12], SwinT [19]) and hybrid models (SegFormer+ViT, SegFormer+SwinT, SwinT+ViT), and our proposed model. Table 3 depicts that SwinT achieved mIoU of 64.1% and mDice of 82.7%, while SegFormerB5 performed slightly lower with 60.4% mIoU, 70.7% mDice. DeepLabv3+ underperformed in comparison. Among hybrid models, SwinT+ViT demonstrated improved performance (mIoU: 73.08%, mDice: 84.62%), outperforming SegFormer+ViT (mIoU5: 5.89%, mDice: 66.49%) and SegFormer+SwinT (mIoU: 60.93%, mDice: 77.83%). However, our proposed model, which combines the SwinT encoder with the SegFormer decoder, achieved the highest segmentation accuracy with an mIoU of 75.2% and mDice of 85.4%.

In class-wise performance, SwinSegFormer excelled in detecting fine structures like vehicles (IoU: 55.2, Dice: 68.6) and roads (IoU: 82.5, Dice: 90.4), while hybrid models with SegFormer encoders struggled, especially with classes like vehicles (IoU: 19.72, Dice: 32.94 for SegFormer+SwinT). The SwinT+ViT model performed competitively (mIoU: 73.08, mDice: 84.62) but still lagged in boundary refinement compared to SwinSegFormer (mIoU: 75.2, mDice: 85.4). These results confirm that the SwinT [19] hierarchical encoder effectively captures local and global contextual information, while the SegFormer MLP decoder refines the segmentation masks. These experiments depict SwinSegFormer an effective hybrid model that addresses local and global features in flood-related imagery and improves results for challenging classes in the FloodNet dataset [42], resulting in a superior performance with comparatively fewer parameters, as shown in Table 4.

2) QUANTITATIVE ANALYSIS

Ablation research confirms SwinSegFormer's superior performance in flood image segmentation over other standalone and hybrid models. SwinSegFormer consistently achieves higher scores in metrics with a 75.2% mIOU, 85.4% mDice, 85.4% mFscore, reported in Table 4. This robust performance across various measures highlights its capability to capture local and global information crucial for flood detection. Our experiments on the FloodNet dataset, focusing on challenging classes like flooded roads, non-flooded roads, and vehicles, showcase SwinSegFormer's ability. Fine-tuning SwinT, SegFormer and other hybrid models like SegFormer+ViT, SegFormer+SwinT, SwinT+ViT on FloodNet dataset demonstrates their capabilities, but SwinSegFormer consistently outperforms them. Specifically, in key classes, SwinSegFormer attains a 55.29% IOU and 68.67% dice score for the vehicle class, 63.53% IOU and 77.69% dice score for the flooded road class, 65.58% IOU and 79.21% dice score for the pool class, and 82.45% IOU and 90.44% dice score for non-flooded roads (as shown in Table 3) which outperforms SwinT, SegFormer, hybrid models (SegFormer+ViT, SegFormer+SwinT, and SwinT+ViT.

3) QUALITATIVE ANALYSIS

The SwinSegFormer model outperforms competitors both quantitatively and qualitatively. Visual analysis consistently

TABLE 4. Performance and Computational Resources Comparison of Proposed Model With SOTA

Methods	mIOU	mDice	mFscore	Encoder Parameters	Decoder Parameters	Overall Parameters
DeepLabV3+ [41]	50.4%	65.4%	65.4%	42.5M	20.1M (MLP)	62.6M
SegFormerB2 [12]	49.0%	60.5%	60.5%	24.2M	3.3M (MLP)	27.5M
SegFormerB4 [12]	50.7%	65.0%	65.0%	60.8M	3.3M (MLP)	64.1M
SegFormerB5 [12]	60.4%	75.2%	70.7%	81.4M	3.3M (MLP)	84.7M
SwinT [19]	64.1%	82.7%	75.2%	88.0M	≈ 29.7M (UperNet SwinT)	≈ 117.7M
SegFormer + ViT	55.9%	16.5%	66.5%	81.4M	≈ 27M (UperNet ViT)	≈ 108.4M
SegFormer + SwinT	60.9%	77.8%	77.8%	81.4M	≈ 29.7M (UperNet SwinT)	≈ 111.1M
SwinT + ViT	73.1%	84.6%	84.2%	88.0M	≈ 27M (UperNet ViT)	≈ 115M
SwinSegFormer	75.2%	85.4%	85.4%	88.0M	3.3M (MLP)	91.3M

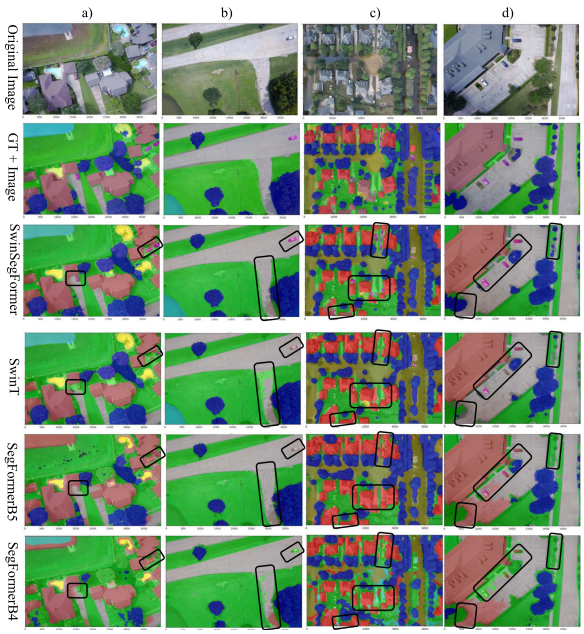

FIGURE 7. Qualitative compariosn of transformer-based models, i.e., SegFormerB4 [12], SegFormerB5 [12], SwinT [19], and SwinSegFormer. Black rectangles highlights that proposed model accurately segments the narrow masked classes such as vehicles, roads, and buildings both flooded (F) and non-flooded (NF).

delivers more precise and detailed semantic segmentation results for flooded and non-flooded areas. This superiority is especially evident in challenging scenarios involving small objects like vehicles, pools, and complex classes like flooded and non-flooded roads. SwinSegFormer excels in accurately segmenting cars and pools, while other models struggle to do so and provide a better tree mask (as shown in Fig. 7). Compared to other models, SwinSegFormer shows fewer misclassifications and smoother boundaries. Its ability to capture fine details in remotely sensed images makes it a compelling choice for flood detection applications. These visual results align with quantitative findings, affirming SwinSegFormer's comprehensive improvement in visual fidelity and accuracy over other assessed models.

E. COMPARATIVE STUDY

The SwinSegFormer model outperforms several SOTA models in flood detection (shown in Table 5), establishing itself

TABLE 5. mIOU Comparison of SwinSegFormer With SOTA Using FloodNet

Networks	mIOU(%)
UNet-MobileNetV3 [41]	59.3
DeepLabV3+ [28]	61.5
UNetFormer [44]	47.2
HardNet-70 [46]	57.9
SegFormerB0 [12]	61.6
SwinSegFormer (Proposed Model)	75.2

as an innovative breakthrough in flood detection. It excels in a comprehensive comparative study, as demonstrated by various evaluation metrics, including mIOU, mDice, and F1score (Table 6), and depicts that the proposed model surpasses benchmark models. Additional metrics of mACC, mPrecision, and mRecall provided in Section VI.

Our visualization study further supports these results, showing SwinSegFormer's superior object boundary delineation, class discrimination, and reduced misclassifications compared to competitors. The model's accuracy in flood detection is evident in the segmentation results, where it excels in segmenting narrow masked classes, vehicles, pools, roads, flooded buildings, and challenging vehicle classes. Fig. 7 shows our model's superior performance, particularly in accurately segmenting narrow-masked classes. Fig 7(a) shows that our model outperforms SOTA models by accurately segmenting vehicles and non-flooded roads (shown with black rectangles). Fig. 7(b) demonstrate the model's ability to accurately segment narrow masked classes like roads and vehicles, as highlighted with black rectangles. Additionally, Fig. 7(c) displays the better segmentation results obtained for flooded buildings and the narrow-masked vehicle class in the FloodNet dataset. The black rectangles indicate small vehicles that were efficiently segmented. Fig. 7(d) shows our model's exceptional performance in segmenting vehicles with clear boundaries, achieving efficient tree segmentation, and accurately labeling grass and roads, surpassing other models. In conclusion, SwinSegFormer emerges as leading choice for precise flood detection, offering valuable insights and improved disaster response capabilities.

TABLE 6. Comparing IOU, Dice, and F1 Scores of Our SwinSegFormer Model With SegFormer [12] Variants,i.e., B2, B4, B5 and Other SOTA Models

	IOU						Dice						F1score					
Classes	B2	B4	B5	[19]	[12]	our	B2	B4	B5	[19]	[12]	our	B2	B4	B5	[19]	[12]	our
Building(F)	58.5	57.6	68.6	71.8	67.3	77.7	73.8	73.1	81.4	83.6	80.4	87.4	73.8	73.1	81.34	83.5	80.4	87.4
Building(NF)	43.0	51.6	62.8	77.7	56.9	78.6	60.2	68.1	77.2	87.4	72.5	88.1	60.1	68.1	77.1	87.4	72.5	88.0
Road(F)	33.8	42.6	50.3	58.7	23.5	63.5	50.6	59.8	66.9	74.0	38.1	77.6	50.5	59.7	66.9	73.9	38.1	77.6
Road(NF)	59.4	69.3	71.7	78.3	70.1	82.5	74.5	81.8	83.5	87.9	82.4	90.4	74.5	81.8	83.5	87.8	82.4	90.4
Water	44.8	50.7	60.4	67.6	60.5	72.1	61.9	67.3	75.3	80.7	75.4	83.7	61.8	67.2	75.3	80.6	75.4	83.7
Tree	72.7	73.4	75.5	73.6	49.6	85.1	84.2	84.7	86.0	84.8	66.3	91.9	84.1	84.6	86.0	84.7	66.3	91.9
Vehicle	2.3	4.6	18.6	30.9	9.5	55.2	4.6	8.8	31.3	47.2	17.5	68.6	4.5	8.7	31.3	47.2	17.5	68.6
Pool	26.7	26.2	37.5	49.6	41.3	65.5	42.2	41.5	54.5	66.3	58.4	79.2	42.1	41.4	54.5	66.2	58.4	79.2
Grass	81.3	80.9	83.0	85.3	75.9	89.1	89.7	89.4	90.7	92.1	86.3	94.2	89.6	89.4	90.6	92.0	86.3	94.2
Mean	49.0	50.7	60.4	64.1	50.4	75.2	60.5	65.0	75.2	82.7	65.4	85.4	60.4	65.0	70.6	75.1	65.4	85.4

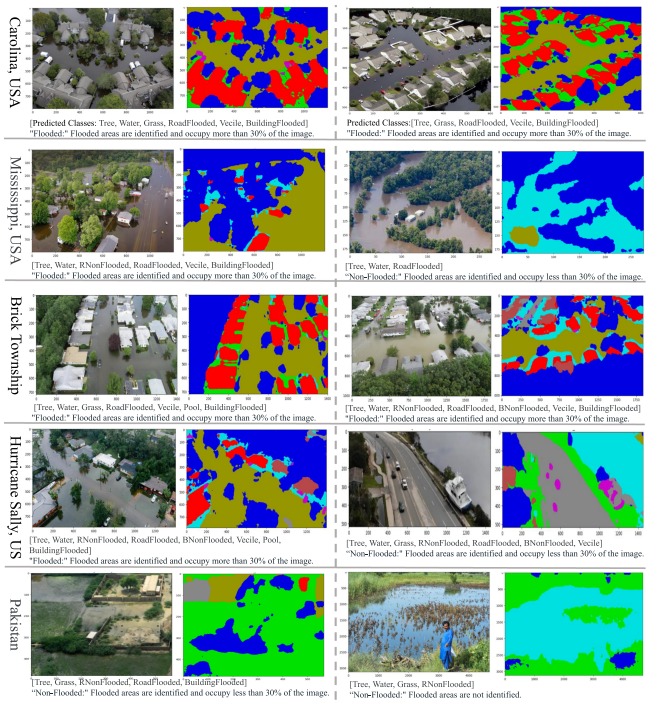

FIGURE 8. Evaluation of SwinSegFormer by testing with real-world images. Our model effectively segments classes, estimate flood-affected areas, and classifies them into flooded (F) and non-flooded (NF).

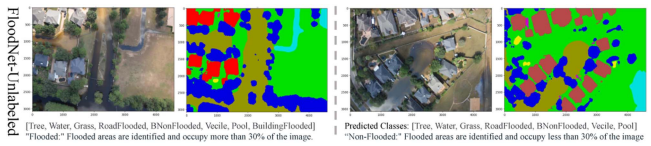

FIGURE 9. Evaluating SwinSegFormer with unlabel FloodNet data: shows that our model accurately predicts classes, estimate flood-affected areas, and classifies them into flooded and non-flooded.

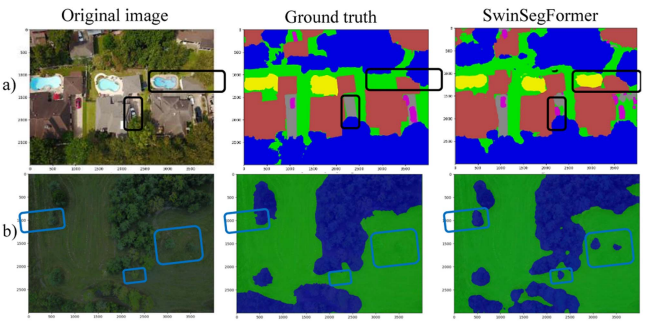

FIGURE 10. Comparison between ground truth and model predictions highlights the challenges and limitations of FloodNet dataset. Please observe the black and blue rectangles, where our model able to segments minor details but lacks corresponding ground truths.

1) ESTIMATION OF FLOODED-AFFECTED AREA IN REAL-WORLD IMAGES

We have provided the semantic segmentation results for real-world flood-affected images using SwinSegFormer. In Fig. 8, we included images from five different flood disasters in each row, including record-breaking rainfall in Charleston, South Carolina, USA, flood along the Mississippi River in USA, coastal storms in Brick Township, Hurricane Sally leading to flash floods in the USA, and floods in Pakistan. Our post-processing module displays all predicted classes and calculates the flood-affected area for each segmented result. Image classifiesd as flooded have a flood-affected area of more than 30%, while non-flooded images have less.

2) ESTIMATION OF FLOODED AREA WITH UNLABELED FLOODNET DATA

We employed the proposed trained model to analyze unlabeled images from FloodNet dataset. The GitHub repository includes approximately 2343 labeled and 1047 unlabeled images, which we are using to make inferences. Fig. 9 depicts the the model's efficiency in segmenting unlabeled images. These findings collectively underscore the effectiveness of our proposed model. Subsequently, the post-processing module calculates the extent of the flood-affected area and classifies it into flooded and non-flooded regions.

F. CHALLENGES IN FLOODNET DATASET

As mentioned earlier, we employed FloodNet dataset for this research and discovered significant limitations of FloodNet dataset that impacted the performance of the model. Specifically, our model successfully segmented certain true classes that were missed in the ground-truth annotations, leading to a perceived decline in overall performance. Many images contain narrow masks or small objects that are often overlooked during annotation. As an example, we have included Fig. 10(a) and (b) to illustrate these challenges. Black rectangles in Fig. 10(a) depict classes such as pools and vehicles that are not included in the annotations. Similarly, Fig. 10(b)

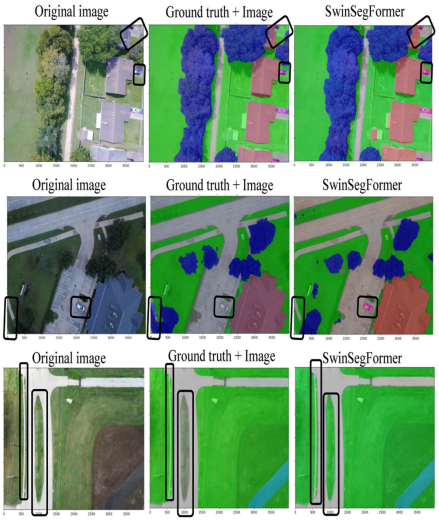

FIGURE 11. Black rectangle shows unlabel/mislabel road, grass and vehicle class in ground truth.

TABLE 7. Performance Comparison of Proposed Model With SOTA

Methods	mRecall	mPrecision	mACC
DeepLabV3+ [41]	62.5%	85.1%	62.6%
SegFormerB2 [12]	61.9%	65.5%	60.5%
SegFormerB4 [12]	64.8%	67.6%	65.0%
SegFormerB5 [12]	70.4%	74.1%	71.8%
SwinT [19]	79.1%	77.0%	75.2%
SegFormer + ViT	65.7%	79.3%	66.7%
SegFormer + SwinT	76.6%	77.4%	76.6%
SwinT + ViT	85.5%	84.5%	83.5%
SwinSegFormer	86.7%	85.1%	85.4%

shows that small trees with narrow masks (highlighted in blue rectangles) were also ignored. This reduces the performance of the model for the corresponding classes. Despite these challenges, proposed SwinSegFormer model was able to accurately segment these ignored classes. To further improve the performance, ground-truth can be re-annotated in the future, leading to substantial improvements in the model's performance.

V. CONCLUSION

This work presented a comprehensive research on the detection of flood-affected areas using aerial imagery, focusing on harnessing the potential of vision transformers in remote sensing. We utilized FloodNet dataset, containing high-resolution UAV images captured after Hurricane Harvey, to develop an efficient model for segmenting flooded areas. Our proposed SwinSegFormer model demonstrates that the SwinSegFormer model excels in segmenting the challenging classes such as vehicles, pools, flooded roads, and non-flooded roads, outperforming the state-of-the-art models. Moreover, the performance of the proposed model is verified by making inferences over real-world flooded images.

This capability enables the rescue teams to understand the extent of damage in an area and formulate the rescue plans accordingly. This extends the application of vision transformers in remote sensing and flood detection application, offering promising capabilities for disaster management. Results demonstrate that the SwinSegFormer's precision in identifying flood-affected areas with affordable computational resources can improve the flood monitoring, decision-making, and emergency response strategies in vulnerable regions, efficiently. While the SwinSegFormer demonstrates promising performance on the FloodNet dataset, future research can explore its application in diverse flood scenarios across different environmental conditions and regions. Enhancing the model's ability to handle varying data quality and annotation styles from different sources would increase its adaptability. Future work could also focus on improving the model's robustness in challenging scenarios, such as those involving occlusions or visually ambiguous regions. These advancements would further improves the model's performance for real-world flood detection and disaster management applications.

ADDITIONAL RESULTS

In our article, we have already provided comprehensive experimental results. This section presents supplementary findings that demonstrate the superior performance of our proposed model. Table 7 presented additional performance metrics, i.e., recall, precision, and accuracy to compare the performance of our model with SOTA models including DeeplabV3+ [41], SegFormerB2/B4/B5 [12], SwinT [19] and hybrid models (SegFormer + ViT, SegFormer + SwinT, SwinT + ViT), evidencing improved performance.

Moreover, Fig. 11 illustrates annotation inconsistencies in the dataset (highlighted with black rectangles), particularly in narrow regions such as roads. While this discrepancy impacts model performance, our architecture demonstrates robust segmentation capabilities. Fig. 11(a) depicts the input image, while Fig. 11(b) and (c) showcase the corresponding ground-truth and our model's segmentation output, respectively, highlighting accurate segmentation of fine-grained classes. This observation underscores the disparity in ground-truth annotations and identifies a potential gap for dataset refinement in future research endeavors.

REFERENCES

[1] S. Jiang, C. Jiang, and W. Jiang, "Efficient structure from motion for large-scale UAV images: A review and a comparison of SFM tools," *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 230–251, 2020.
[2] T. Hoeser and C. Kuenzer, "Object detection and image segmentation with deep learning on earth observation data: A review-part I: Evolution and recent trends," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1667.
[3] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.
[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[6] J. Kang, L. Liu, F. Zhang, C. Shen, N. Wang, and L. Shao, "Semantic segmentation model of cotton roots in-situ image based on attention mechanism," *Comput. Electron. Agriculture*, vol. 189, 2021, Art. no. 106370. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169921003872

[7] Y. Li et al., "CTMU-NET: An improved U-Net for semantic segmentation of remote-sensing images based on the combined attention mechanism," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 10148–10161, 2023.

[8] H. Iqbal, D. Campo, P. Marin-Plaza, L. Marcenaro, D. M. Gómez, and C. Regazzoni, "Modeling perception in autonomous vehicles via 3D convolutional representations on LiDAR," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14608–14619, Sep. 2022.

[9] H. Iqbal, A. Al-Kaff, P. Marin, L. Marcenaro, D. M. Gomez, and C. Regazzoni, "Detection of abnormal motion by estimating scene flows of point clouds for autonomous driving," in *Proc. IEEE Int. Intell. Transp. Syst. Conf.*, 2021, pp. 2788–2793.

[10] S. Liu, J. Cheng, L. Liang, H. Bai, and W. Dang, "Light-weight semantic segmentation network for UAV remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8287–8296, 2021.

[11] H. Luo, C. Chen, L. Fang, X. Zhu, and L. Lu, "High-resolution aerial images semantic segmentation using deep fully convolutional network with channel attention mechanism," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3492–3507, Sep. 2019.

[12] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 12077–12090.

[13] W. Shi, M. Zhang, H. Ke, X. Fang, Z. Zhan, and S. Chen, "Landslide recognition by deep convolutional neural network and change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4654–4672, Jun. 2021.

[14] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[15] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Compu. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.

[16] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568–578.

[17] Z. Xiao et al., "CapMatch: Semi-supervised contrastive transformer capsule with feature-based knowledge distillation for human activity recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 2, pp. 2690–2704, Feb. 2025.

[18] Z. Xiao et al., "Densely knowledge-aware network for multivariate time series classification," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 54, no. 4, pp. 2192–2204, Apr. 2024.

[19] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[20] Y. Wang, X. Wang, and J. Jian, "Remote sensing landslide recognition based on convolutional neural network," *Math. Problems Eng.*, vol. 2019, 2019, Art. no. 8389368.

[21] Y. Chong, X. Chen, and S. Pan, "Context union edge network for semantic segmentation of small-scale objects in very high resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2020, Art no. 6000305.

[22] S. Xiang, Q. Xie, and M. Wang, "Semantic segmentation for remote sensing images based on adaptive feature selection network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art no. 8006705.

[23] B. Yu, L. Yang, and F. Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3252–3261, Sep. 2018.

[24] V. Sangeetha and K. Prasad, "Syntheses of novel derivatives of 2-acetylfuro [2, 3-] carbazoles, benzo [1, 2-b]-1, 4-thiazepino [2, 3-] carbazoles and 1-acetyloxycarbazole-2-carbaldehydes," *Indian J. Chemistry Sect. B-Organic Chemistry Including Medicinal Chemistry*, vol. 45, no. 8, pp. 1951–1954, 2006.

[25] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1492–1500.

[26] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2020.

[27] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art no. 4408820.

[28] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[29] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1520–1528.

[30] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[31] W. Weng and X. Zhu, "INet: Convolutional networks for biomedical image segmentation," *IEEE Access*, vol. 9, pp. 16591–16603, 2021.

[32] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted Res-UNet for high-quality retina vessel segmentation," in *Proc. 9th Int. Conf. Inf. Technol. Med. Educ.*, 2018, pp. 327–331.

[33] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected unet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[34] R. Li et al., "DeepUNet: A deep fully convolutional network for pixel-level sea-land segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 3954–3962, Nov. 2018.

[35] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[36] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[37] S. Du and M. Liu, "Class-guidance network based on the pyramid vision transformer for efficient semantic segmentation of high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5578–5589, 2023.

[38] R. J. Chen et al., "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16144–16155.

[39] Y. Xu et al., "FDViT: Improve the hierarchical architecture of vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 5950–5960.

[40] Z. Xue, X. Tan, X. Yu, B. Liu, A. Yu, and P. Zhang, "Deep hierarchical vision transformer for hyperspectral and LiDAR data classification," *IEEE Trans. Image Process.*, vol. 31, pp. 3095–3110, 2022.

[41] F. Safavi, T. Chowdhury, and M. Rahnemoonfar, "Comparative study between real-time and non-real-time segmentation models on flooding events," in *Proc. IEEE Int. Conf. Big Data*, 2021, pp. 4199–4207.

[42] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy, "FloodNet: A high resolution aerial imagery dataset for post flood scene understanding," *IEEE Access*, vol. 9, pp. 89644–89654, 2021.

[43] F. Safavi and M. Rahnemoonfar, "Comparative study of real-time semantic segmentation networks in aerial images during flooding events," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 15–31, 2022.

[44] L. Wang et al., "UNetFormer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 190, pp. 196–214, 2022.

[45] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[46] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 325–341.

MUHAMMAD TARIQ SHAHEEN received the B.Sc. degree in 2020 and the master's degree in electrical engineering from the National University of Science and Technology, Islamabad, Pakistan. His research interests include image processing, computer vision, machine learning, deep learning, and dense prediction tasks using vision transformer-based techniques in the area of geoscience and remote sensing.

HAFSA IQBAL received the International Joint Doctorate degree *(cum laude)* from the University of Genoa, Genoa, Italy, and Carlos III University of Madrid, Madrid, Spain, respectively. She is currently a Postdoc with the Carlos III University of Madrid, Madrid, Spain. Her research interests include computer vision, machine learning and deep learning techniques for cognitive and interactive environments. She was the recipient of the President's Gold Medal during her M.S. degree.

NUMAN KHURSHID received the Ph.D. degree in AI from the Lahore University of Management Sciences (LUMS), Lahore, Pakistan. He is currently an Assistant Professor with the Schools of Electrical Engineering and Computer Sciences, National University of Sciences and Technology, Islamabad, Pakistan. His research interests include deep learning and generative AI for remote sensing signals.

HALEEMA SADIA received the Ph.D. degree in electrical engineering in 2024. From 2021 to 2024, she was a Graduate Research Assistant with GIK Institute, Khyber Pakhtunkhwa, Pakistan. She is currently a Postdoctoral Research Fellow with the Department of Electrical and Communication Engineering, College of Engineering, United Arab Emirates University, Al Ain, UAE. Her research interests include wireless communication networks, UAVs, optimization techniques for B5G networks, and ML for 5G.

NASIR SAEED (Senior Member, IEEE) received the B.Sc. degree in telecommunication from the University of Engineering and Technology, Peshawar, Pakistan, in 2009, and the M.Sc. degree in satellite navigation from Polito di Torino, Turin, Italy, in 2012, and the Ph.D. degree in electronics and communication engineering from Hanyang University, Seoul, South Korea, in 2015. From 2015 to 2017, he was an Assistant Professor with the Department of Electrical Engineering, IQRA National University, Peshawar. From July 2017 to December 2020, he was a Postdoctoral Research Fellow with the Communication Theory Laboratory, King Abdullah University of Science and Technology (KAUST), Saudi Arabia. He is currently an Associate Professor with the Department of Electrical and Communication Engineering at United Arab Emirates University (UAEU), Al Ain, UAE. He has authored or coauthored more than 80 international journal and conference articles. His research interests include non-conventional communication networks, heterogeneous vertical networks, multi-dimensional signal processing, and localization. He is also an Associate Editor for IEEE WIRELESS COMMUNICATIONS LETTERS.