# Hybrid Vision Transformers and CNNs for Enhanced Transmission Line Segmentation in Aerial Images

Hoanh Nguyen*, Tuan Anh Nguyen
Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam

***Abstract*—This paper presents a novel architecture for the segmentation of transmission lines in aerial images, utilizing a hybrid model that combines the strengths of Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs). The proposed method first employs a Swin Transformer backbone (Swin-B) that processes the input image through a hierarchical structure, effectively capturing multi-scale contextual information. Following this, an upsampling strategy is employed, wherein the features extracted by the transformer are refined through convolutional layers, ensuring that the resolution is maintained, and spatial details are recovered. To integrate multi-level feature maps, a feature fusion module with a squeeze-and-excitation (SE) layer is introduced, which consolidates the benefits of both high-level and low-level feature extractions. The SE layer plays a pivotal role in augmenting the feature channels, focusing the model's attention on the most informative features for transmission line detection. By leveraging the global receptive field of ViTs for comprehensive context and the local precision of CNNs for fine-grained detail, our method aims to set a new benchmark for transmission line segmentation in aerial imagery. The effectiveness of our approach is demonstrated through extensive experiments and comparisons with existing state-of-the-art methods.**

***Keywords*—Vision transformers; convolutional neural networks; transmission lines segmentation; hybrid model; feature fusion**

## I. INTRODUCTION

Transmission line segmentation in aerial images is a critical task in the maintenance and monitoring of electrical power grids. It enables the automated inspection of power lines for fault detection, vegetation encroachment, and structural analysis, which are essential for ensuring the reliability and safety of electricity distribution. However, this task is fraught with challenges. Aerial images often have highly variable lighting conditions, weather effects, and diverse landscapes that can obscure the visibility of transmission lines. Additionally, the lines themselves can be difficult to distinguish due to their thin and linear nature against complex backgrounds. Another significant challenge is the presence of other linear structures, such as roads and railways, that can be easily confused with power lines by automated systems. The movement of the aerial platform, whether it's a drone or a manned aircraft, introduces motion blur and varying angles of capture, further complicating the segmentation process. Addressing these challenges requires robust algorithms capable of high precision and adaptability to a range of environmental conditions and image qualities. Traditional methods for vision-based transmission line detection and segmentation have revolved around the utilization of edge and line segment detection techniques as foundational steps. These methods generate a multitude of potential cable segment candidates by applying algorithms such as the Hough transform, Radon transform, and various other heuristic and search-based line detection strategies, such as the circle-based search, heuristic line detection, Line Segment Detector (LSD), and Edge Drawing for line segment detection (EDLines). Once edges and line segments are detected, a set of specialized rules, informed by the structural characteristics of cables and the context of their surroundings, are applied to discern correct cable segments and eliminate false positives. One of the main drawbacks of these approaches is their dependency on numerous parameters and complex rules that need to be meticulously set by hand, which hinders their adaptability to different environments. As a result, the precision and robustness of traditional methods can be significantly compromised due to environmental variations, making it challenging to maintain consistent performance across diverse scenarios. The advent of deep learning has catalyzed significant advancements in the domain of vision-based transmission line detection and segmentation. CNN-based methods have eclipsed traditional techniques, demonstrating substantial improvements in detection accuracy and computational efficiency. These deep neural networks facilitate end-to-end learning and inference, simplifying the complex parameter tuning process inherent in multistage approaches and enhancing generalizability across varied scenarios. For instance, CNNs have been trained to identify image patches containing cables, which are then further processed using traditional methods like the Hough transform for line segmentation. Moreover, some approaches have integrated fully convolutional networks with line segment regressors for direct line segment detection, particularly effective in scenarios where aerial images capture transmission lines at close range. However, when dealing with long-range and wide-angle captures that result in indistinct or slightly curved cable representations, these methods pivot towards a pixel-wise segmentation framework, employing semantic and instance segmentation techniques to provide a more nuanced cable detection and enable individual cable instance identification, which is pivotal for autonomous UAV applications.

Recent years have witnessed the rapid development of ViTs [1, 2]. ViTs leverages the transformer architecture, originally designed for natural language processing, to handle sequences of image patches as input. ViTs model relationships between these patches through self-attention mechanisms, making them capable of capturing global dependencies within an image. The combination of CNNs and ViT into a hybrid architecture aims to harness the local feature extraction proficiency of CNNs with the global context understanding of ViTs. CNNs are adept at recognizing patterns and textures within small regions of an image, making them excellent for tasks that require detailed local information, such as edge detection. ViTs, with their attention-based approach, can consider the entire image at once, which allows for a more holistic understanding of the scene. By integrating both, the hybrid model can effectively process and integrate both local and global information, leading to improved performance on complex tasks like transmission line segmentation. This synergy can provide a more nuanced understanding of images, enabling the model to be both precise in detail and comprehensive in scope, potentially overcoming limitations found in models that rely on a single approach.

This study introduces an innovative hybrid architecture for the precise segmentation of transmission lines in aerial images, leveraging the synergistic potentials of ViTs and CNNs. The core of our proposed method is a Swin Transformer backbone, adept at hierarchically processing the input image to encapsulate multi-scale contextual information. This is complemented by an upsampling mechanism that meticulously refines the transformer-extracted features via convolutional layers, crucial for preserving resolution and restoring spatial details. A feature fusion module, equipped with an SE layer, is integrated to merge feature maps from multiple levels, harnessing both the high-level and low-level extraction strengths. The SE layer is instrumental in enhancing feature channels, directing the model's focus towards the most salient features for detecting transmission lines. Our approach is designed to exploit the expansive receptive field of ViTs for global context awareness, while utilizing the CNNs' local precision for capturing intricate details, thereby establishing a new standard for transmission line segmentation in aerial photography. The method's superiority is validated through comprehensive experimental benchmarks, showcasing its advancement over current leading methodologies.

The rest of the paper is organized as follows: Section II presents related studies; Section III details our proposed model; Section IV describes the experiments and results; Section V provides the conclusions.

## II. Related Work

### A. Transmission Line Detection and Segmentation

Traditional approaches to vision-based detection and segmentation of transmission lines have primarily focused on employing edge and line segment detection techniques as their fundamental processes. In study [3], a real-time algorithm was developed for detecting power lines in UAV video images, where the process begins with converting video images into binary images using adaptive thresholding. Subsequently, Hough Transform identifies line candidates in these binary images, and a fuzzy C-means clustering algorithm discriminates actual power lines from these candidates. Mu et al. [4] proposed a method for automatically extracting power lines from cluttered natural backgrounds in aerial images. The approach involves using a Gabor filter to eliminate background noise, followed by the application of the Hough transform to detect straight lines in the images. Zhang et al. [5] introduced a new method for detecting and tracking power lines, starting with the use of the Hough transform to extract line segments. The method then employs K-means clustering in the Hough space to filter and identify power lines and utilizes a Kalman filter for tracking these lines within the continuity of a video sequence. In study [6], the authors presented an algorithm that capitalizes on the geometric relationships inherent to circle symmetry for line segment detection. It employs Canny and Steerable Filters to detect line segments, which are then linked in a subsequent stage for effective analysis. Sharma et al. [7] introduced a novel morphological operator and robust image space heuristics for the accurate location and complete extraction of power lines. Santos et al. [8] introduced PLineD, a new vision-based power line detection algorithm designed to robustly detect power lines, even in noisy image backgrounds. Although traditional approaches to vision-based detection and segmentation of transmission lines have achieved some success, they still have many limitations. A significant limitation of these methods is their reliance on a multitude of parameters and intricate rules that require careful manual adjustment, impeding their flexibility across various environments. Consequently, the accuracy and reliability of traditional approaches are often adversely affected by environmental changes, posing challenges in achieving uniform performance in different settings. With the outstanding advantages of CNNs, many methods using CNNs for transmission line segmentation have been proposed [9]. In [10], the authors introduced a pyramidal patch classification framework that effectively eliminates clutter without relying on additional auxiliary tools. This is achieved through a hierarchical patch partition and selection strategy, complemented by a new spatial grid pooling layer in the CNN-based classifier. Nguyen et al. [11] presented LS-Net, a rapid, single-shot line-segment detector tailored for power line detection, which is fully convolutional by design and comprises three modules: a fully convolutional feature extractor, a classifier, and a line segment regressor. Lee et al. [12] presented a weakly supervised learning algorithm for identifying power lines. The algorithm classifies sub-regions within images using a sliding window approach and a CNN. In [13], a Transmission Line Detection (TLD) algorithm, CableNet, is proposed, drawing inspiration from instance segmentation and incorporating enhancements to Fully Convolutional Networks (FCNs) [14] with overlaying dilated and spatial convolutional layers for better representation of transmission lines, and dual output branches for generating multidimensional feature maps for instance segmentation.

### B. Vision Transformer-CNN Hybrid Models

In recent years, the rapid development of ViTs has significantly advanced the field of computer vision, leading to their widespread application in tasks ranging from image classification to complex scene understanding [15, 16]. Instead of simplifying ViTs, another prominent research direction involves merging components of ViTs and CNNs to create novel backbone architectures. These hybrid models combine the local feature extraction prowess of CNNs with the global contextual understanding afforded by ViTs, thus offering a comprehensive approach to image analysis. Follow this approach, [17] highlighted the adaptation of principles from the extensive literature on CNNs, particularly the use of activation maps with decreasing resolutions, to enhance the design of transformers. The study in [18] investigated the optimization challenges of ViT models, attributing the issues to their 'patchify stem' design, and proposes a solution by replacing it with stacked stride-two 3x3 convolutions. This modification significantly enhances optimization stability and model accuracy, leading to the recommendation of using a standard, lightweight convolutional stem in ViT models for improved performance and robustness. In study [19], the authors introduced BoTNet, a versatile and efficient backbone architecture for various computer vision tasks, which enhances performance by integrating self-attention mechanisms. This is achieved by replacing spatial convolutions with global self-attention in the last three bottleneck blocks of a ResNet, leading to notable improvements in instance segmentation and object detection, while simultaneously reducing the number of parameters and maintaining minimal latency overhead. ConViT [20] presented the concept of gated positional self-attention (GPSA), a novel form of positional self-attention designed with a flexible, 'soft' convolutional inductive bias. GPSA layers are initially configured to emulate the locality characteristic of convolutional layers but are also equipped with a gating parameter that allows each attention head to dynamically balance the focus between positional and content information. Guo et al. [21] proposed a novel hybrid network that synergizes the long-range dependency capturing capabilities of transformers with the local information extraction prowess of CNNs. Recently, PVTv1 [22], PVTv2 [23], LITv1 [24], and LITv2 [25] incorporate convolutional operations at each stage of ViT models to diminish the token count and construct hybrid, multi-stage structures.

## III. METHOD

### A. *Model Architecture*

Fig. 1 illustrates the overall pipeline of our method, which integrates a vision transformer encoder with a convolutional neural network decoder to create a hybrid model for the segmentation of transmission lines in aerial images. Input images undergo a hierarchical processing through multiple layers of the Swin Transformer [26], each reducing the spatial dimensions while increasing the depth of feature representation. These layers (Layer 1 to Layer 4) progressively transform the input, capturing intricate details and contextual information at various scales. The transformed features are then upsampled and passed through convolutional layers to refine the feature maps, ensuring that spatial information is preserved and enhanced. The upsampling process gradually restores the resolution of the feature maps, which are then combined through feature fusion steps. These fusion steps are essential as it aggregates multi-scale information, enabling the model to capture both high-level semantic information and low-resolution spatial details. A squeeze-and-excitation (SE) layer is subsequently employed to recalibrate the feature channels, emphasizing informative features while suppressing less useful ones. Finally, a segmentation head, comprising a series of convolutional layers, is responsible for generating the output segmentation map that delineates the transmission lines within the aerial images.

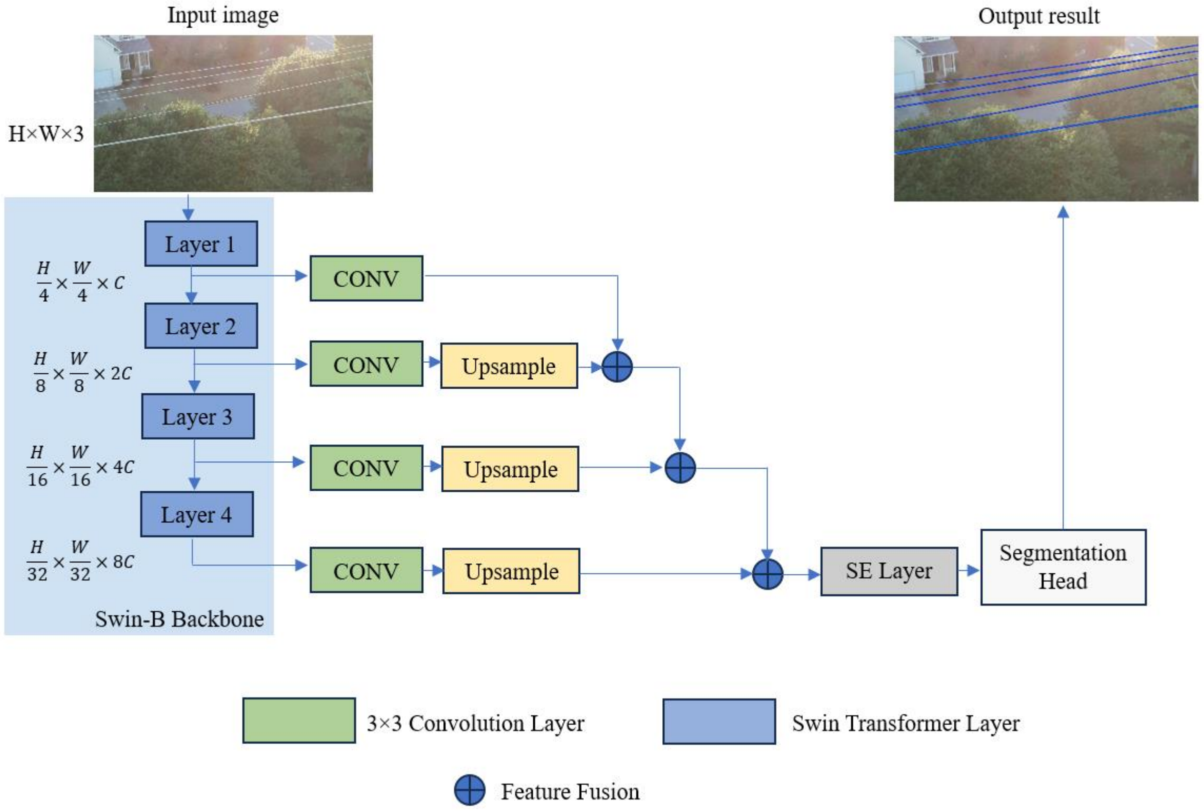

Fig. 1. Model architecture.

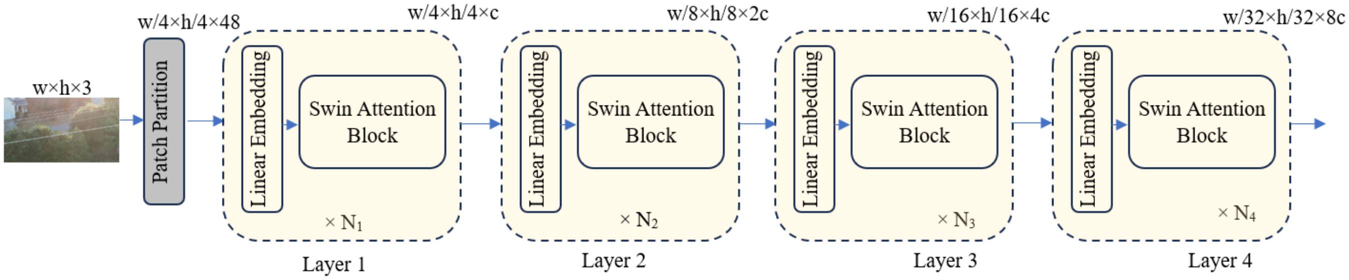

Fig. 2. The structure of Swin Transformer encoder.

TABLE I. DETAILED ARCHITECTURE OF SWIN-B

| Layer | Output Size | Number of Blocks | Attention Heads | Attention Head Dimensions | MLP Ratio |
|---|---|---|---|---|---|
| Patch Partition | w/4 × h/4 × 48 | N/A | N/A | N/A | N/A |
| Layer 1 | w/4 × h/4 × 96 | 2 | 3 | 32 | 4 |
| Layer 2 | w/8 × h/8 × 192 | 2 | 6 | 32 | 4 |
| Layer 3 | w/16 × h/16 × 384 | 6 | 12 | 32 | 4 |
| Layer 4 | w/32 × h/32 × 768 | 2 | 24 | 32 | 4 |

### B. Swin-B-based Encoder

Fig. 2 illustrates the structure of Swin Transformer encoder. The original input image is represented as *w×h×3*, where *w* and *h* are the width and height of the image, and *3* represents the RGB color channels. This image is then partitioned into patches. The size of these patches is 4×4 pixels, which are then flattened and linearly embedded into a higher-dimensional space (e.g., 48 features per patch). So, the input dimension to the first transformer layer is *w/4×h/4×48*. As the input passes through each Swin Transformer layer consisting consecutive $N_i$ ($i$ = 1, 2, 3, 4) Swin transformer blocks, the spatial resolution is further reduced, and the feature dimensionality is increased, thus enhancing the model's ability to capture more complex features at different scales. The Swin Transformer employs a self-attention mechanism within each block. The attention is computed using queries ($Q$), keys ($K$), and values ($V$), which are derived from the input feature maps. The formulation of self-attention within the Swin Transformer involves a sequence of operations beginning with the computation of $Q$, $K$, and $V$ via linear transformations of the input feature map. The attention scores are then determined by calculating the dot product between $Q$ and $K$. These scores are normalized using a softmax function to derive attention weights, which are subsequently used to obtain a weighted feature representation by multiplying them with $V$. To ensure stability in the gradients during training, a scaling factor, commonly the inverse square root of the keys' dimensionality, is optionally applied to the dot product of $Q$ and $K$. Mathematically, the attention can be represented as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where, $d_k$ is the dimensionality of the key vectors. This allows the model to focus on different parts of the image depending on the learned importance of each feature.

The Swin Transformer framework is offered in four distinct variants, namely Swin-T (Tiny), Swin-S (Small), Swin-B (Base), and Swin-L (Large), each differing in capacity and computational requirements. The choice of a particular Swin Transformer variant for a given task hinges on a balance between the model's empirical performance and the computational constraints of the available hardware. In the case of the segmentation of transmission lines in aerial images, Swin-B was selected due to its robust performance in capturing intricate details and providing a higher feature resolution necessary for the precise delineation of transmission lines, which are often slender and require fine-grained detection capabilities. Moreover, Swin-B strikes a balance between computational efficiency and model complexity, making it a pragmatic choice for tasks demanding high accuracy without exceedingly intensive computational demands. Table I provides detailed architecture of Swin-B backbone.

### C. CNNs-based Decoder

The decoder leverages a U-Net-like structure known for its effectiveness in segmentation tasks due to its ability to combine low-level feature maps with high-level ones, thus capturing context and fine details. Each feature map output from the Swin-B backbone passes through a 3×3 convolution layer. This operation serves to refine the feature maps by applying filters that can capture spatial hierarchies within the data. After the convolution layers, the feature maps are upsampled. This process increases the spatial resolution of the feature maps to prepare them for feature fusion. The upsampling doubles the height and width of the feature maps, as is common in U-Net architectures [27] to match the dimensions of the feature maps from the encoder that will be fused. The upsampled feature maps are then fused with corresponding feature maps from earlier layers of the encoder. This step is crucial as it reintroduces higher resolution details that may have been lost during downsampling in the encoder. Feature fusion is done using element-wise addition operation. The last part of the decoder is the segmentation head, which outputs the final segmentation map. This head takes the processed feature maps and applies a combination of convolutional layers, activation functions, and sigmoid layer to generate the pixel-wise classification of the transmission lines. Given the size of the input to the decoder is $\frac{H}{32} \times \frac{W}{32} \times 8C$, the size of the output should match the original height and width of the input image $H \times W \times 2$.

### D. Squeeze-and-Excitation Layer

Before the final output, we employ a Squeeze-and-Excitation (SE) layer [28] to recalibrate the feature channels by explicitly modelling the interdependencies between them. The SE layer uses global average pooling to squeeze global spatial information into a channel descriptor, then uses two fully connected layers to capture channel-wise dependencies, and finally applies the channel weights back to the original feature maps to emphasize useful features and suppress less useful ones. Fig. 3 shows the architecture of the SE layer. Let the input to the SE layer be a feature map $F$ with dimensions $H' \times W' \times C$, where $H'$ and $W'$ are the spatial dimensions after upsampling and $C$ is the number of channels. The SE layer first performs a global average pooling operation on $F$, which squeezes the spatial dimensions $H'$ and $W'$ into a single channel descriptor $z$ with dimensions $1 \times 1 \times C$ . Mathematically, this is represented as:

$$z_c = \frac{1}{H' \times W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} F_{i,j,c} \qquad (2)$$

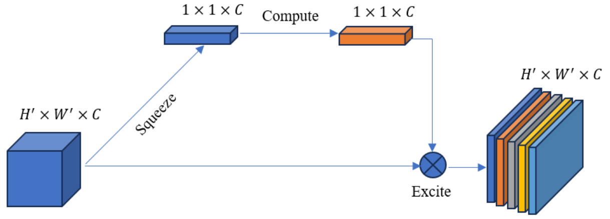

Fig. 3. The architecture of the SE layer.

where, $z_c$ is the *c-th* element of $z$; $F_{i,j,c}$ is the value at position *(i, j)* in channel *c* of the feature map *F*.

The SE layer then passes *z* through two fully connected layers. The first layer reduces the channel dimensionality from $C$ to $\frac{C}{r}$ using a ReLU activation function, and the second layer increases it back to $C$ using a sigmoid activation function, thus generating the channel-wise weights *s* with the same dimension $1 \times 1 \times C$. Mathematically, this is represented as:

$$s = \sigma\big(g(z, W)\big) = \sigma(W_2.\theta(W_1.z)) \qquad (3)$$

where, $\sigma$ denotes the sigmoid activation, $\theta$ denotes the ReLU activation, $W_1$ and $W_2$ are the weights of the fully connected layers, and $g$ represents the excitation function.

Finally, the SE layer applies these weights *s* back to the original feature map *F* through channel-wise multiplication, producing the output feature map $F'$ with the same spatial dimensions $H' \times W'$ but with recalibrated channels:

$$F'_{i,j,c} = s_c.F_{i,j,c} \qquad (4)$$

This operation scales each channel of the input feature map by the corresponding learned weight, emphasizing informative features and suppressing less relevant ones. The output of the SE layer is then ready to be passed to the subsequent layers in the decoder for further processing towards the final segmentation map.

### E. Loss Function

Binary Cross-Entropy (BCE) loss is a commonly used loss function for binary classification tasks, such as the segmentation of transmission lines in aerial images, where each pixel is classified as either belonging to a transmission line (positive class) or background (negative class). The BCE loss function measures the distance between the predicted probabilities and the actual binary labels, penalizing predictions that diverge from the true labels. Formally, the BCE loss for a single pixel is calculated as:

$$L_i = -[ylog(p) + (1-y)log(1-p)] \qquad (5)$$

where, $y$ is the true label of the pixel, and $p$ is the predicted probability that the pixel belongs to the transmission line class. The true label $y$ is 1 if the pixel is part of a transmission line and 0 otherwise. The predicted probability $p$ is obtained from the output of a sigmoid activation function in the last layer of the neural network, ensuring that $p$ is in the range [0,1].

For the entire image, the total BCE loss is the average of the individual pixel losses:

$$L_{BCE} = -\frac{1}{N} \sum_{i=1}^{N} L_i \qquad (6)$$

where, $N$ is the total number of pixels in the image.

## IV. Experiments

### A. Dataset and Metrics

We use the TTPLA dataset [29] to evaluate the proposed method. The TTPLA dataset is a specialized collection of aerial images designed for the detection and segmentation of transmission towers and power lines. This dataset is significant for training and evaluating machine learning models, particularly in the domain of remote sensing and automated monitoring of electrical infrastructure. It includes high-resolution images that capture the intricate details of transmission towers and power lines from various angles and under different lighting conditions. The diversity of the dataset aids in developing robust models capable of accurately identifying and segmenting these structures. The dataset consists of 1,100 images with a resolution of 3,840×2,160 pixels and manually labeled 8,987 instances of transmission lines and transmission towers. For the purpose of evaluating the transmission line segmentation task, we only employ the labels of transmission lines for training and testing.

Precision (*P*), recall (*R*), Intersection over Union (*IoU*), and *F-score* are critical metrics for evaluating the performance of models in the segmentation of transmission lines in aerial images. Precision measures the ratio of correctly predicted positive observations to the total predicted positives. It is formulated as:

$$P = \frac{TP}{TP+FP} \qquad (7)$$

where, $TP$ is true positives and $FP$ is false positives.

Recall assesses the ratio of correctly predicted positive observations to all actual positives. It's given by:

$$R = \frac{TP}{TP+FN} \tag{8}$$

with $FN$ being false negatives.

Intersection over Union (*IoU*), also known as the Jaccard index, is the area of overlap between the predicted segmentation and the ground truth divided by the area of union. The formula is:

$$IoU = \frac{\text{area of overlap}}{\text{area of union}} \tag{9}$$

*F-score* is the harmonic mean of precision and recall, providing a balance between them. It's calculated by:

$$F - score = 2 \times \frac{P \times R}{P+R} \tag{10}$$

These metrics are pivotal for tuning models to the specific challenges of aerial image segmentation, such as delineating thin and often indistinct transmission lines against complex backgrounds. High precision indicates a model that reliably identifies line pixels, while high recall shows it finds most of the actual line pixels. *IoU* gives an overall sense of the model's accuracy, and *F-score* offers a single measure to assess both precision and recall.

### B. Implementation Details

We leverage the strengths of transformer models, specifically building upon the Swin Transformer for the encoder component. The Swin-B model is pretrained on ImageNet-22k with a resolution of 384×384, maintaining the window size (M) as in the pretrained models. To adapt to the higher resolution requirements of the semantic segmentation task, we fine-tune these models based on the dataset's resolution. Following methodologies in the literature, relative position bias is incorporated when calculating attention scores. The decoders are initialized with random weights from a normal distribution. For optimization, the AdamW optimizer [30] is employed. The input aerial images are resized to a standard resolution of 512×512 pixels, striking a balance between preserving detail and maintaining computational efficiency. The learning rate is set at 1e-4, paired with a cosine annealing scheduler to facilitate adaptive learning rate adjustments over approximately 100 training epochs. The model's training is executed on a high-end NVIDIA RTX 4080 GPU. A batch size of 4 is used. Implementation is carried out using deep learning frameworks PyTorch. To further bolster the model's ability to generalize, data augmentation techniques including random rotations, flipping, scaling, and brightness adjustments are employed.

### C. Comparison with Existing Methods on TTPLA Dataset

We compared the proposed model with six existing models on the TTPLA Dataset, as shown in Table II. It is evident that our model, which integrates a vision transformer encoder with a convolutional neural network decoder, outperforms most of the existing models in terms of precision, IoU (Intersection over Union), and F-score. These metrics are critical for assessing the effectiveness of segmentation models in aerial imagery. Notably, the proposed model achieves a precision of 0.855 and an F-score of 0.671, surpassing the UNet, UNet++, and Focal-UNet models that also use the Resnet-18 architecture. This superior performance can be attributed to the efficient feature representation and fusion enabled by the hybrid architecture of the proposed model. The Swin Transformer layers in the encoder capture intricate details and contextual information at various scales, which is crucial for accurately delineating transmission lines in complex aerial images. The use of the squeeze-and-excitation layer further enhances the model by emphasizing informative features, allowing for a more nuanced segmentation output. This is evident in the comparative improvement in precision and F-score, where the model excels in correctly identifying relevant pixels while maintaining high overall segmentation accuracy. In contrast, models like LCNN and HAWP, based on the Hourglass architecture, show significantly lower precision and F-score values, indicating a lesser ability to accurately segment transmission lines in aerial images. Their lower performance might be due to less effective feature extraction and fusion compared to the proposed hybrid model. Overall, the proposed model's superior performance across multiple metrics, especially in precision and F-score, highlights its effectiveness in segmenting transmission lines in aerial images, demonstrating the advantages of its novel architecture combining vision transformer and convolutional layers.

TABLE II. SEGMENTATION PERFORMANCE OF THE PROPOSED METHOD AND THE COMPARISON METHODS ON THE TTPLA DATASET

| Models | Backbone | P | R | IoU | F-score |
|---|---|---|---|---|---|
| DeepLabv3+ [31] | Resnet-18 | 0.784 | 0.510 | 0.424 | 0.573 |
| UNet [27] | Resnet-18 | 0.846 | 0.583 | 0.515 | 0.662 |
| UNet++ [32] | Resnet-18 | 0.843 | 0.591 | 0.522 | 0.668 |
| Focal-UNet [33] | Resnet-18 | 0.784 | 0.577 | 0.504 | 0.662 |
| LCNN [34] | Hourglass | 0.541 | 0.315 | 0.498 | 0.519 |
| HAWP [35] | Hourglass | 0.581 | 0.421 | 0.485 | 0.532 |
| Our Model | Swin-B | 0.855 | 0.579 | 0.522 | 0.671 |

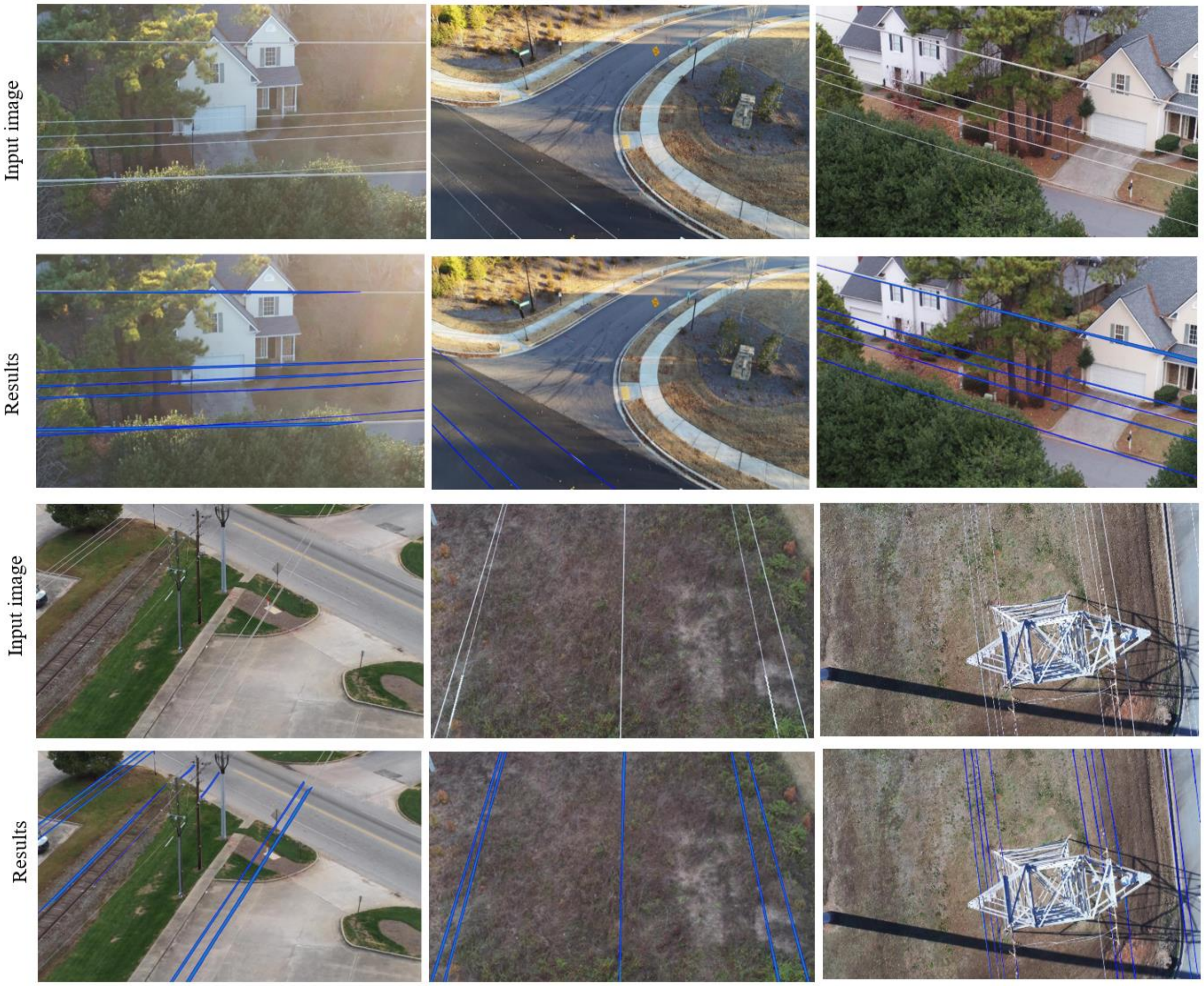

Fig. 4. Sample transmission line segmentation results of the proposed model.

Fig. 4 displays a side-by-side comparison of original aerial images against the segmentation results produced by the proposed model for transmission line identification. Across various landscapes such as residential areas, road intersections, and open fields with transmission towers, the model delineates the transmission lines with a high degree of precision, as indicated by the blue lines overlaying the images. Specifically, the model accurately identifies transmission lines in residential areas without confusion from similar-colored backgrounds, showing its precision in challenging environments. At a road intersection, the model successfully differentiates transmission lines from road markings despite visual noise, indicating strong feature extraction capabilities. In open fields, the model effectively handles contrasts and textures, maintaining accurate segmentation over uniform backgrounds like grass. These results underline the model's robustness in varied settings and its applicability in real-world tasks such as infrastructure monitoring from aerial imagery.

### *D. Comparison of Swin-B with Other State-of-the-Art Backbones*

We conducted experiments to evaluate the performance of Swin-B encoder. Fig. 5 shows the F-score performance of various backbone architectures employed in image segmentation models. Swin-B tops the chart with an F-score of 0.671, indicating its superior ability in combining features effectively for precise segmentation tasks. Swin-T closely trails with a marginally lower F-score of 0.668, suggesting that while it is slightly less effective than Swin-B, it remains a highly competitive architecture. ResNeSt-101 and ResNet-101, both advanced iterations of the ResNet family, score 0.642 and 0.630 respectively, pointing to a proficient but noticeably lesser segmentation capability compared to the Swin architectures. VGG-16, the oldest architecture among those compared, shows its limitations with an F-score of 0.585, underscoring the advancements in backbone architectures for segmentation tasks and the importance of choosing the right one for optimal performance.

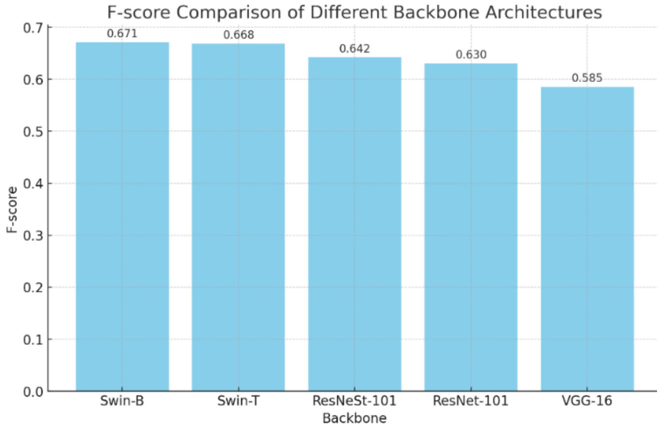

Fig. 5. F-score comparison of different backbone architectures.

## V. Conclusion

This paper presents a novel hypbid architecture for the segmentation of transmission lines in aerial images. The presented hybrid segmentation model, which leverages the synergy of a vision transformer encoder and a convolutional neural network decoder, has proven highly effective in the segmentation of transmission lines from aerial images. The model's performance, as demonstrated on the TTPLA Dataset, is superior to existing models, achieving remarkable precision and F-score metrics. Its ability to handle complex backgrounds and maintain high accuracy in diverse environments showcases its robustness and adaptability. The successful application of this model paves the way for its integration into aerial survey systems, offering significant improvements in the monitoring and maintenance of power line infrastructures, potentially reducing costs and increasing operational efficiency. The research outcomes not only contribute to the advancement of segmentation techniques but also underline the transformative impact of integrating transformer architectures within computer vision tasks. For future work, we will incorporate advanced object detection algorithms to identify not just transmission lines but also associated structures such as towers and insulators. This would provide a more comprehensive analysis of the aerial imagery, facilitating detailed inspections and maintenance planning.

## References

[1] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent.*, Virtual Event, Austria, May 2021, pp. 1–22.

[2] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[3] Yang, Tang Wen, Hang Yin, Qiu Qi Ruan, Jian Da Han, Jun Tong Qi, Qing Yong, Zi Tong Wang, and Zeng Qi Sun. "Overhead power line detection from UAV video images." In *2012 19th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, pp. 74-79. IEEE, 2012.

[4] Mu, Chao, Jie Yu, Yanming Feng, and Jinhai Cai. "Power lines extraction from aerial images based on Gabor filter." In *International Symposium on Spatial Analysis, Spatial-Temporal Data Modeling, and Data Mining*, vol. 7492, pp. 1081-1088. SPIE, 2009.

[5] Zhang, Jingjing, Liang Liu, Binhai Wang, Xiguang Chen, Qian Wang, and Tianru Zheng. "High speed automatic power line detection and tracking for a UAV-based inspection." In *2012 International Conference on Industrial Control and Electronics Engineering*, pp. 266-269. IEEE, 2012.

[6] Cerón, Alexander, and Flavio Prieto. "Power line detection using a circle based search with UAV images." In *2014 international conference on unmanned aircraft systems (ICUAS)*, pp. 632-639. IEEE, 2014.

[7] Sharma, Hrishikesh, Rajeev Bhujade, V. Adithya, and P. Balamuralidhar. "Vision-based detection of power distribution lines in complex remote surroundings." In *2014 Twentieth National Conference on Communications (NCC)*, pp. 1-6. IEEE, 2014.

[8] Santos, Tiago, Miguel Moreira, J. Almeida, André Dias, Alfredo Martins, J. Dinis, J. Formiga, and E. Silva. "PLineD: Vision-based power lines detection for Unmanned Aerial Vehicles." In *2017 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pp. 253-259. IEEE, 2017.

[9] Jenssen, Robert, and Davide Roverso. "Automatic autonomous vision-based power line inspection: A review of current status and the potential role of deep learning." *International Journal of Electrical Power & Energy Systems* 99 (2018): 107-120.

[10] Li, Yan, Chaofeng Pan, Xianbin Cao, and Dapeng Wu. "Power line detection by pyramidal patch classification." *IEEE Transactions on Emerging Topics in Computational Intelligence* 3, no. 6 (2018): 416-426.

[11] Nguyen, Van Nhan, Robert Jenssen, and Davide Roverso. "LS-Net: Fast single-shot line-segment detector." *Machine Vision and Applications* 32 (2021).

[12] Lee, Sang Jun, Jong Pil Yun, Hyeyeon Choi, Wookyong Kwon, Gyogwon Koo, and Sang Woo Kim. "Weakly supervised learning with convolutional neural networks for power line localization." In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1-8. IEEE, 2017.

[13] Li, Bo, Cheng Chen, Shiwen Dong, and Junfeng Qiao. "Transmission line detection in aerial images: An instance segmentation approach based on multitask neural networks." *Signal Processing: Image Communication* 96 (2021): 116278.

[14] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431-3440. 2015.

[15] Alwajih, Fakhraddin, Eman Badr, and Sherif Abdou. "Transformer-based Models for Arabic Online Handwriting Recognition." *International Journal of Advanced Computer Science and Applications* 13, no. 5 (2022).

[16] Gutiérrez Choque, Anyelo Carlos, Vivian Medina Mamani, Eveling Castro Gutiérrez, Rosa Núñez Pacheco, and José Ignacio Aguaded. "Transformer based Model for Coherence Evaluation of Scientific Abstracts: Second Fine-tuned BERT." (2022).

[17] Graham, Benjamin, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. "Levit: a vision transformer in convnet's clothing for faster inference." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12259-12269. 2021.

[18] Xiao, Tete, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. "Early convolutions help transformers see better." *Advances in neural information processing systems* 34 (2021): 30392-30400.

[19] Srinivas, Aravind, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. "Bottleneck transformers for visual recognition." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16519-16529. 2021.

[20] d'Ascoli, Stéphane, Hugo Touvron, Matthew L. Leavitt, Ari S. Morcos, Giulio Biroli, and Levent Sagun. "Convit: Improving vision transformers with soft convolutional inductive biases." In *International Conference on Machine Learning*, pp. 2286-2296. PMLR, 2021.

[21] Guo, Jianyuan, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. "Cmt: Convolutional neural networks meet vision transformers." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12175-12185. 2022.

[22] Wang, Wenhai, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 568-578. 2021.

[23] Wang, Wenhai, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. "Pvt v2: Improved baselines with pyramid vision transformer." *Computational Visual Media* 8, no. 3 (2022): 415-424.

[24] Pan, Zizheng, Bohan Zhuang, Haoyu He, Jing Liu, and Jianfei Cai. "Less is more: Pay less attention in vision transformers." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, pp. 2035-2043. 2022.

[25] Pan, Zizheng, Jianfei Cai, and Bohan Zhuang. "Fast vision transformers with hilo attention." *Advances in Neural Information Processing Systems* 35 (2022): 14541-14554.

[26] Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012-10022. 2021.

[27] Ronneberger, Olaf, Philipp Fischer, and Thomas 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*, pp. 234-241. Springer International Publishing, 2015.

[28] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132-7141. 2018.

[29] Abdelfattah, Rabab, Xiaofeng Wang, and Song Wang. "Ttpla: An aerial-image dataset for detection and segmentation of transmission towers and power lines." In *Proceedings of the Asian Conference on Computer Vision*. 2020.

[30] Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." *arXiv preprint arXiv:1711.05101* (2017).

[31] Chen, Liang-Chieh, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation." In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801-818. 2018.

[32] Zhou, Zongwei, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. "Unet++: A nested u-net architecture for medical image segmentation." In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pp. 3-11. Springer International Publishing, 2018.

[33] Jaffari, Rabeea, Manzoor Ahmed Hashmani, and Constantino Carlos Reyes-Aldasoro. "A novel focal phi loss for power line segmentation with auxiliary classifier U-Net." *Sensors* 21, no. 8 (2021): 2803.

[34] Zhou, Yichao, Haozhi Qi, and Yi Ma. "End-to-end wireframe parsing." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 962-971. 2019.

[35] Xue, Nan, Tianfu Wu, Song Bai, Fudong Wang, Gui-Song Xia, Liangpei Zhang, and Philip HS Torr. "Holistically-attracted wireframe parsing." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2788-2797. 2020.