ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume X-5/W2-2025
Unleashing the power of Geospatial & Frontier Technologies for a Sustainable Future – GIFTS Summit 2025,
1–3 September 2025, Symbiosis International University, Pune, India

# Classifying Roof Types from Orthophotos: A Swin Transformer-Based Approach

Shailja[a,*], Manohar Yadav[a], Ramji Dwivedi[a], Poonam Jayhind Pardeshi[a]

aGeographic Information System (GIS) Cell, Motilal Nehru National Institute of Technology Allahabad, Prayagraj-211004, India-
{shailja.2021rgi, ssmyadav, ramjid, poonam.2024rgi01}@mnnit.ac.in

**Keywords:** Rooftop types, Orthophotos, Classification, Transformer.

**Abstract**

Rooftop type classification refers to the process of identifying and categorizing the structural geometry of building roofs using geospatial data. It plays a critical role in urban analysis, aiding applications such as 3D city modeling, solar potential estimation, infrastructure planning and post-disaster damage assessment. This research proposes a transformer-based deep learning model, the Swin Transformer, for rooftop classification to handle the issues of complex roof shapes, small or similar-looking structures, and variations in roof types, especially in areas with unplanned construction. The model is trained on an orthophoto-derived GeoTIFF images having four roof type such as flat, gable, complex and bug. Images were resized to 256×256 pixels and processed in batches of 128. This dataset is split into 2528 training images, 544 testing images, and 545 validation images. The Transformer architecture achieves overall performance with a test accuracy of 75%, showing excellent results for gable classes having F1-score of 85.23% and complex classes achieving F1-score of 70.75%, while flat and bugs classes show moderate performance due to lower recall. With the integration of early stopping and a learning rate scheduler, the Swin Transformer showed improved precision for bugs from 60.94% to 66.67% and flat from 78.41% to 66.93% classes, while maintaining a comparable overall accuracy 75.00% to 74.63% and enhancing class balance in predictions. The proposed architecture is also compared with other state-of-the art models and can be used in future applications such as distinguishing roofs from other urban and roof improved geospatial analysis and smart city development.

## 1. Introduction

Building rooftop type classification refers to the process of identifying and categorizing different roof shapes using data from aerial or satellite imagery, LiDAR, or other remote sensing sources. In the era of rapid urbanization and smart city initiatives, accurate classification of rooftop types has gained increasing importance for a variety of urban applications such as 3D city modeling, solar energy potential estimation, infrastructure management, urban planning and climate resilience planning (Biljecki et al., 2015). Conventional approaches for rooftop classification have primarily relied on hand-crafted features, digital surface models (DSM) and geometric cues such as lines, planes, and ridges (Li et al., 2022). While methods based on DSM data and stereo matching offer height-based segmentation, they often suffer from noise, occlusion, and inaccuracies at object boundaries. To address these challenges, recent research has shifted towards deep learning (DL) methods, especially Convolutional Neural Networks (CNNs) (O'Shea & Nash, 2015) and their variants. CNNs have demonstrated high accuracy in image-based classification tasks but require extensive labeled datasets and often struggle with class imbalance and generalization to diverse urban settings.

Researchers have explored CNNs and machine learning models for rooftop classification. Mohajeri et al. (2018) applied an SVM-based model to classify 10,085 rooftops in Geneva by shape for solar suitability, achieving 66% accuracy, with flat and shed roofs showing the highest PV potential. Castagno & Atkins (2018) utilized CNN models like ResNet-50 and Inception on RGB and LiDAR data from cities such as Manhattan and Witten, reaching up to 88.3% accuracy. Buyukdemircioglu et al. (2021) developed a CNN pipeline for classifying six roof types from high-resolution orthophotos in Turkey, obtaining accuracy rates above 80% for most classes. Comparative evaluations by (Partovi et al., 2017)highlighted the complementary roles of CNN and SVM in handling different roof geometries from WorldView-2 data. Yildirim & Karsli (2024) proposed a deep learning architecture using 3D point cloud data for roof classification on the RoofN3D dataset. (Alidoost & Arefi, 2019) introduced a CNN-based two-step approach using a single aerial image for building detection and roof type classification. (Wang et al., 2022) addressed rural roof classification from UAV imagery using an improved Mask R-CNN with ResNet152 and visual feature fusion (e.g., VDVI and Sobel edge maps).

However, a major limitation of these models is their dependency on large, annotated datasets and their sensitivity to visual similarity between roof types (e.g., flat vs. bug-shaped). This limits their scalability in heterogeneous urban environments, especially in developing regions with unplanned constructions. Advent in deep learning technologies, particularly Transformer-based architectures (Vaswani, 2017), offer a promising direction in the field of image classification, segmentation and object detection. Transformers have outperformed CNNs in many computer vision tasks by utilizing self-attention mechanisms to handle long-range spatial dependencies and capture the global context of an image, whereas CNNs primarily focus on local features.

This study proposes a transformer based, Swin Transformer architecture (Liu et al., 2021) for rooftop type classification from high-resolution GeoTIFF orthophotos. The Swin Transformer can work well even with smaller labeled datasets because it uses pre-trained weights and processes image features more efficiently through its layered attention design. Unlike previous CNN models, this approach handles small objects, irregular geometries, and noise in densely built environments with better precision. The model was trained to classify rooftops into four categories such as flat, gable, complex, and bug. In addition to the Swin Transformer, several state-of-the-art models based on both CNN and Transformer-based architectures were implemented to provide a comparative evaluation. All models were tested under two experimental

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-5-W2-2025-577-2025  

conditions, with and without training callbacks such as early stopping and learning rate schedulers to assess their impact on overall performance and class-level precision and recall. Special focus was given to addressing class imbalance, which commonly affects real-world geospatial datasets, by analyzing the improvements in minority class predictions under different training setups. In rooftop classification, many rooftops were found to appear visually similar or to overlap in orthophotos, which often led to overfitting and unstable learning. To address this, the Swin Transformer was evaluated both with and without callbacks, to determine whether any performance changes were driven by the model's in-built ability to capture fine geospatial patterns or by the stabilizing effects of training optimizations. The rest of the paper comprises of Section 2 which describes the dataset and methodology, Section 3 that presents the results and discussion and Section 4 provides the conclusion with future research.

## 2. Method

This section explains the dataset employed for the building rooftop type classification study and explains the processing steps undertaken. Following that, the classification models employed in the study are explained. Subsequently, the experimental configuration is presented, along with its performance for this research.

### 2.1 Dataset

The dataset used in this study comprises 3,617 GeoTIFF images categorized into four classes: flat, gable, complex, and bugs. These images were cropped with a 2-meter buffer around rooftops and masked accordingly. The "bugs" class contains samples that are not valid rooftop such as construction areas, blurry or partial images, non-roof or unclear structures. The source for these images is a high-resolution orthophoto captured in 2020 using an UltraCam Eagle Mark 3 aerial camera. The flight altitude during image capture was of the range 2,850-3,200 meters, with 60% longitudinal and 30% transverse overlap between images. The total area covered was approximately 1,961 sq. km, focusing on Sofia's Metropolitan Municipality (1,342 sq. km). For this project, the Lozenets district was used as the study area, spanning 9.2 sq. km. The images have a 10 cm/pixel resolution and include RGBA bands across 39 georeferenced image tiles.(Hristov et al., 2023).

The dataset used in this study consists of four classes are labeled as Bugs**,** Complex_data**,** Flat_data**,** and Gable_hip_other. All rooftop images were resized to 256×256 pixels and normalized by scaling pixel values between 0 and 1. The dataset was divided into training, validation, and testing images with 2528, 545, and 544 images respectively, across four roof classes as shown in Figure 1. To improve model performance and reduce overfitting, data augmentation was applied to the training images using random rotations, zooms, and horizontal or vertical flips. The validation and test sets were not augmented. A one-hot encoding format was used for class labels, and the class distribution was checked, revealing an imbalance. To handle this, class weights were calculated to ensure that each class had a balanced influence during training.

### 2.2 Swin Transformer

In this study, Swin Transformer architecture was implemented for building rooftop type classification objective. The Swin Transformer, a hierarchical vision transformer, introduces a shift-window mechanism that balances local feature extraction and global context, making it well-suited for dense classification tasks in remote sensing as shown in Figure 2. This mechanism efficiently captures both local and global contextual features, which is critical for distinguishing rooftops with similar appearances or irregular shapes. The model takes orthophotos of size 256×256 pixels as input and divides them into small non-overlapping patches. Each patch is embedded into a vector representation that allows the network to capture meaningful features. These features are then passed through a series of Swin Transformer blocks, which apply self-attention within local windows and shifted windows to effectively learn both fine-grained local details and broader spatial patterns. This helps the model distinguish between visually similar roof types and handle complex geometries. After feature extraction, a patch merging layer reduces the spatial dimensions, and global average pooling is applied to summarize the information. Finally, a dense output layer with a softmax activation function classifies each image into one of the four predefined rooftop categories. Unlike CNNs limited by local receptive fields, Swin Transformer is well-suited to differentiate complex, heterogeneous rooftops especially in urban and rural scenes with diverse structures and subtle visual variations.

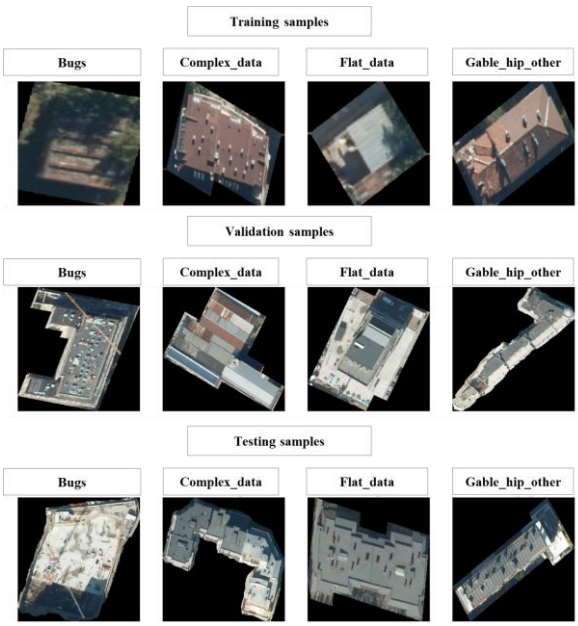



Figure 1. Training, Validation and Testing samples of all the four classes.

### 2.3 Comparison models

To evaluate the performance of the proposed Swin Transformer architecture for rooftop type classification, several state-of-the-art deep learning models were implemented. These models are categorized into transformer-based and convolution-based architectures:

**2.3.1 ViT Classifier**: The Vision Transformer (ViT) introduces a transformer-based approach to image classification, where the input image is split into fixed-size patches, flattened, and embedded into a sequence of tokens (Dosovitskiy et al., 2021). A learnable class token and positional encoding are added, and the sequence is processed through multiple transformer encoder layers. ViT is effective at capturing long-range dependencies and global context but requires a large amount of data to perform optimally. It was included in this study to assess the performance of pure transformer architectures in rooftop classification tasks.

**2.3.2 ConvNeXt**: ConvNeXt is a modernized convolutional neural network inspired by transformer design principles (Liu et al., 2022) which incorporates depthwise convolutions, layer normalization, GELU activations, and a hierarchical architecture, similar to the Swin Transformer. ConvNeXt serves as a strong baseline to compare against transformer models.

**2.3.3 ResNet**: ResNet models are well-known CNN architectures based on residual learning (He et al., 2015). In this study, both ResNet50 (50 layers) and ResNet101 (101 layers) were used. They include skip connections that help in training deeper networks by mitigating vanishing gradient problems. These models are robust for classification tasks and are used to evaluate how traditional CNNs perform against modern attention-based networks. ResNet50 has balanced depth and performance, widely used in vision tasks. ResNet101 has a deeper network allowing more complex feature extraction.

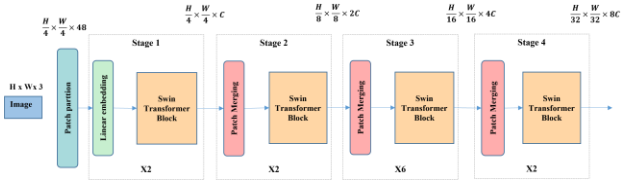



Figure 2. Swin Transformer.

## 2.4 Experimental Configuration

The model is trained for 100 epochs using categorical cross-entropy loss (Rumelhart et al., 1986), and its performance is evaluated on a test set in terms of precision, recall, F1-score (Lipton et al., 2014), and accuracy (Story & Congalton, 1986) across classes. These metrics evaluate the model's classification quality and overall performance. The images are processed in batches of 128 during model training. The model was trained using a class-weighted loss function to address class imbalance, and callbacks like early stopping and learning rate scheduling were applied to improve generalization and optimize training (Prechelt, 1998; Smith, 2017).

# 3. Result & Discussion

This section explains the training, validation and testing performance of Swin Transformer architecture along with comparison models.

## 3.1 Training and Validation performance

To study the effect of training stability, the Swin Transformer model was trained both with and without callbacks. Its performance was then compared with other models such as ViT Classifier, ConvNeXt, ResNet50, and ResNet101. To evaluate training efficiency and avoid overfitting, all models were trained using a maximum of 100 epochs with callbacks applied. The training and validation accuracy and loss curves were obtained. The Swin Transformer completed training at epoch 86, while the ViT Classifier stopped around epoch 81. ConvNeXt converged significantly faster, stopping at epoch 47, and ResNet50 showed early convergence by epoch 14. ResNet101 stopped training at epoch 72. Figures 3, 4, 5, 6 and 7 presents the proposed method and comparison models training and validation performance without and with callbacks. The Swin Transformer showed improved validation performance with callbacks: validation accuracy increased from 72.11% to 74.68%, while training accuracy slightly dropped from 74.01% to 72.67%. Correspondingly, validation loss decreased from 0.6803 to 0.6876, and training loss increased from 0.6544 to 0.6831, indicating better generalization.

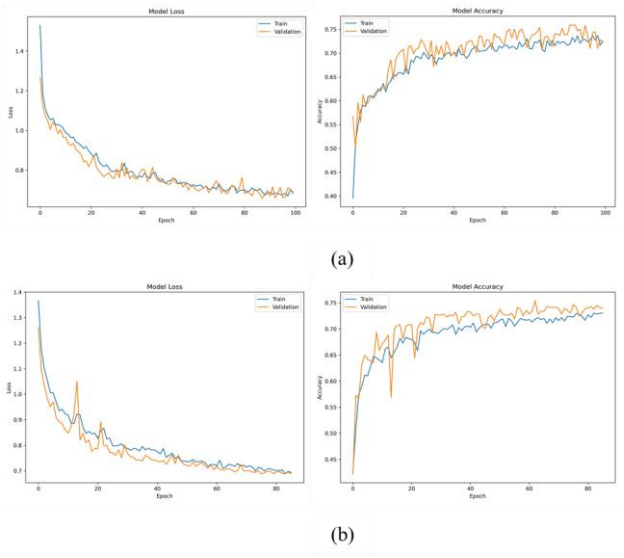



Figure 3. Training and validation loss and accuracy curves for Swin Transformer (a) without and (b) with callbacks.

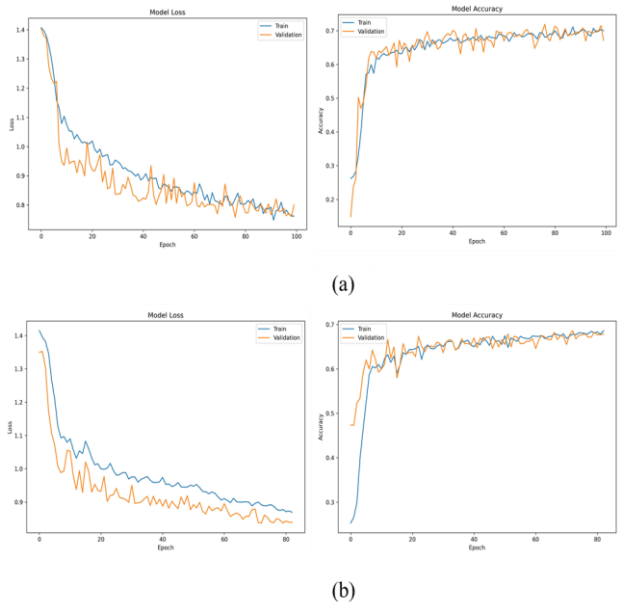



Figure 4. Training and validation loss and accuracy curves for ViT Classifier (a) without and (b) with callbacks.

The ViT Classifier exhibited similar behaviour, where validation accuracy improved from 67.16% to 68.62%, and validation loss decreased from 0.8006 to 0.8369. Training accuracy remained nearly the same (69.38% to 69.58%), while training loss increased slightly from 0.7554 to 0.7998, reflecting a reduced risk of overfitting. The ConvNeXt model benefited significantly from callbacks, with validation accuracy improving from 72.84% to 75.41% and validation loss dropping from 0.6961 to 0.6867. Although training accuracy decreased from 77.53% to 74.72%, the training loss increased from 0.5723 to 0.6405, indicating more stable learning. In contrast, ResNet50 showed decreased performance with callbacks, training accuracy fell from 56.72% to 52.06%, and validation

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-5-W2-2025-577-2025 | © Author(s) 2025. CC BY 4.0 License. 

accuracy from 53.21% to 51.74%. Training loss rose from 1.1278 to 1.2839, and validation loss from 1.1593 to 1.2835. Likewise, ResNet101 experienced a drop in training accuracy from 56.57% to 55.02%, and validation accuracy from 54.86% to 53.76%, with losses increasing from 1.0228 to 1.1864 while training and 1.0567 to 1.1984 at the time of validation.

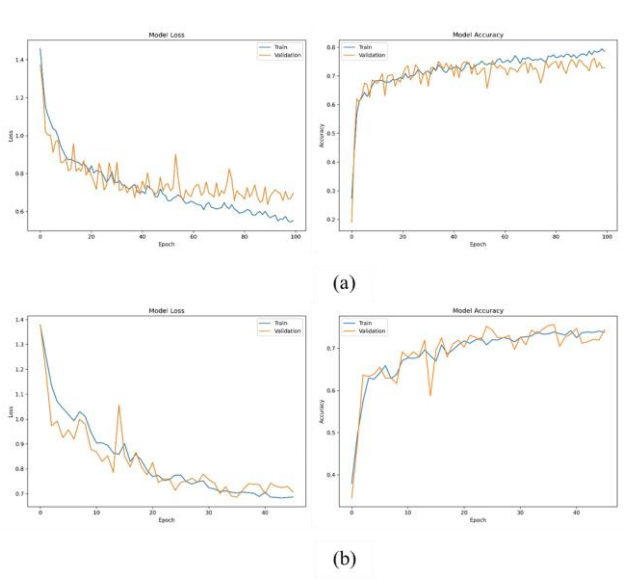



Figure 5. Training and validation loss and accuracy curves for ConvNeXt (a) without and (b) with callbacks.

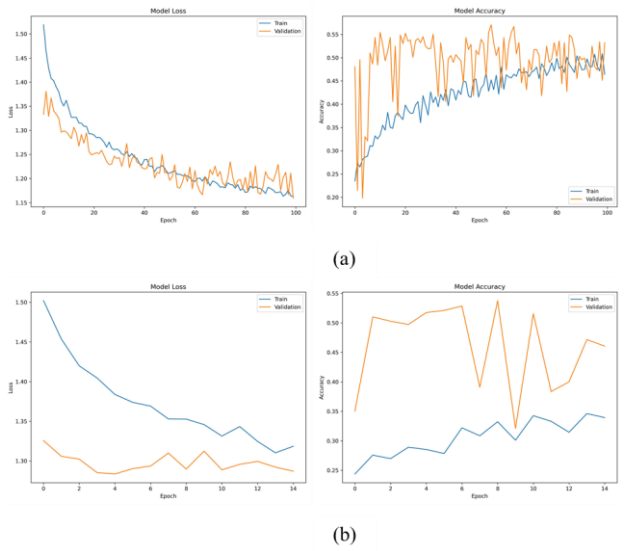



Figure 6. Training and validation loss and accuracy curves for ResNet50 (a) without and (b) with callbacks.

### 3.2 Evaluation of Swin Transformer performance

After training and validation, Swin Transformer was evaluated on a fixed test set comprising 544 georeferenced orthophoto images, evenly representing all four rooftop type classes. To evaluate the classification capabilities of the Swin Transformer model on unseen data, test performance was assessed both with and without callbacks. When trained without callbacks, the model achieved a test accuracy of 75.00%, with a test loss of 0.6935, a macro-average F1-score of 70% and weighted-average F1-score of 74%. Notably, the model performed best on the Gable_hip_other class, achieving a high F1-score of 85%, while performance on the Flat_data and Bugs classes was relatively lower achieving F1-scores of 64% and 59%, respectively, indicating challenges in distinguishing flat rooftops and noisy "bug" samples.

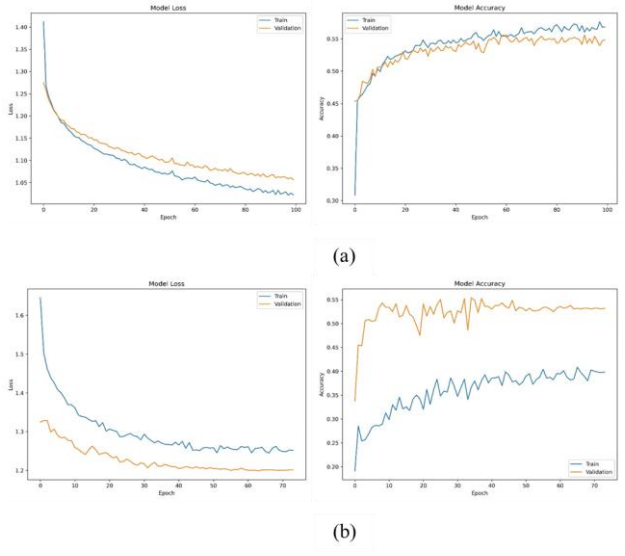



Figure 7. Training and validation loss and accuracy curves for ResNet101 (a) without and (b) with callbacks.

When trained with callbacks, the Swin Transformer achieved a slightly lower test accuracy of 74.63% showing comparable generalization. Flat_data class showed improved F1-score of 67% compared to the non-callback case, while Complex_data and Bugs remained relatively consistent. The Gable_hip_other class retained strong performance attaining F1-score of 85%, indicating that the model is particularly effective at recognizing dominant and geometrically distinct roof types. Table 1 describes the classification report of Swin transformer showing the change in overall accuracy, precision, recall and F1-Score without and with use of callbacks.

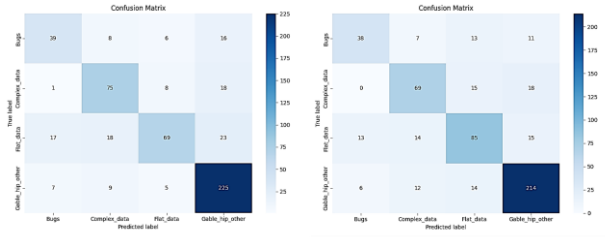
| Class | Without callbacks | | | With callbacks | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1-score | Precision | Recall | F1-score |
| Bugs (69) | 0.61 | 0.57 | 0.59 | 0.67 | 0.55 | 0.60 |
| Complex_data (102) | 0.68 | 0.74 | 0.71 | 0.68 | 0.68 | 0.68 |
| Flat_data (127) | 0.78 | 0.54 | 0.64 | 0.67 | 0.67 | 0.67 |
| Gable_hip_other (246) | 0.80 | 0.91 | 0.85 | 0.83 | 0.87 | 0.85 |
| Overall Accuracy (544) | — | — | 0.75 | — | — | 0.7463 |
| Macro Average (544) | 0.72 | 0.69 | 0.70 | 0.71 | 0.69 | 0.70 |
| Weighted Average (544) | 0.75 | 0.75 | 0.74 | 0.74 | 0.75 | 0.74 |

Table 1. Classification report of Swin Transformer without and with callbacks

Figure 8 presents the confusion matrices for the Swin Transformer model trained without and with callbacks. Without callbacks, the model most accurately classifies the Gable_hip_other class, with 225 correct predictions out of 246. However, there is notable confusion in the Flat_data class, with only 69 correctly predicted and many misclassified as Complex_data (18) and Gable_hip_other (23). The Bugs class

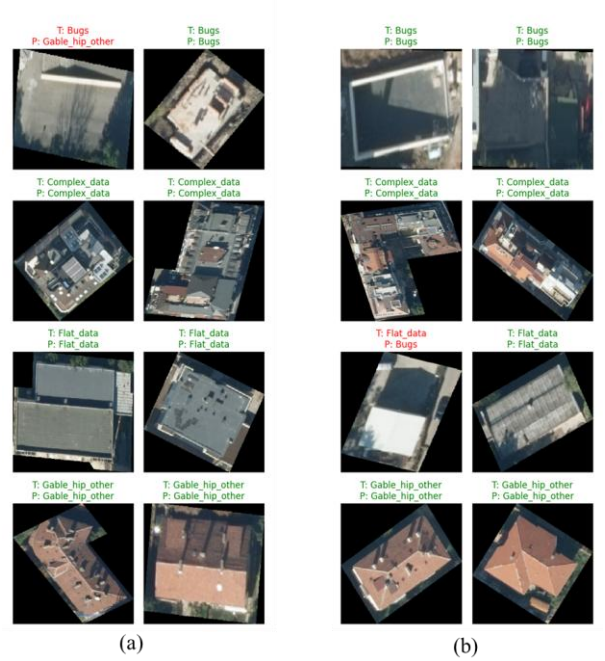
This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-5-W2-2025-577-2025 | © Author(s) 2025. CC BY 4.0 License.

also shows relatively poor performance, with only 39 correct predictions and 16 instances misclassified as Gable_hip_other. When callbacks are applied, the model demonstrates improved balance across all classes. Although Gable_hip_other predictions slightly reduce to 214, the Flat_data predictions increase to 85 correct classifications, and overall misclassifications across classes decline. This reflects improved generalization, showing that callbacks help mitigate overfitting and enhance robustness, particularly for minority and visually overlapping classes.




(a) (b)

Figure 8. Confusion matrix of Swin Transformer (a) without and (b) with callbacks.




(a) (b)

Figure 9. Sample prediction results of the Swin Transformer across all four classes with two test images per class (a) without and (b) with callbacks.

Figure 9 showcases qualitative results of the Swin Transformer model, displaying two sample test images per class. Each row represents one of the four rooftop categories. The columns show predictions under two training scenarios i.e., without callbacks and with callbacks. Each image is represented with the true label (T) and predicted label (P), with correct predictions in green and misclassifications in red. Most images are correctly classified as green labels, but a few misclassifications as red labels highlight the challenges in differentiating between complex or visually similar rooftops. One sample test image of Bugs class was mistakenly predicted as Gable_hip_other when callbacks were not utilized for the training purpose. This misclassification can be attributed to visual overlap in structural patterns, where the small, simple rooftops characteristic of the Bugs class visually resembles extensions or outbuildings often found near Gable or Hip roofs. Whereas in case of callbacks, the model correctly handled the Bugs class but misclassified one Flat_data image as Bugs class. This could be due to flat roofs in small-scale buildings appearing similar in aerial views to isolated bug-like structures, especially if the input lacks context like surrounding structures. The Gable_hip_other class consistently performs well, with both images correctly predicted regardless of callbacks used due to its distinct geometric features. Overall, the Swin Transformer demonstrated strong classification performance across all roof categories, with high accuracy and balanced metrics. The prediction results, confusion matrix, classification report and training and validation curves confirm the ability of Swin Transformer to learn complex patterns and perform well on test data, validating its suitability for rooftop type classification tasks.

### 3.3 Quantitative analysis

This section presents a detailed quantitative evaluation of all models implemented for rooftop type classification in comparison to Swin Transformer. Table 2 presents the classification performance of the ViT Classifier with and without callbacks. Without callbacks, the model achieved an overall accuracy of 67%, which slightly improved to 68% when callbacks were applied.

| Class | Without callbacks | | | With callbacks | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1-score | Precision | Recall | F1-score |
| Bugs (69) | 0.41 | 0.54 | 0.47 | 0.52 | 0.48 | 0.50 |
| Complex_data (102) | 0.51 | 0.57 | 0.54 | 0.44 | 0.65 | 0.52 |
| Flat_data (127) | 0.61 | 0.74 | 0.67 | 0.68 | 0.61 | 0.65 |
| Gable_hip_other (246) | 0.93 | 0.71 | 0.80 | 0.90 | 0.79 | 0.84 |
| Overall Accuracy (544) | — | — | 0.67 | — | — | 0.68 |
| Macro Average (544) | 0.62 | 0.64 | 0.62 | 0.64 | 0.63 | 0.63 |
| Weighted Average (544) | 0.71 | 0.67 | 0.68 | 0.72 | 0.68 | 0.69 |

Table 2. Classification report of ViT Classifier without and with callbacks

| Class | Without callbacks | | | With callbacks | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1-score | Precision | Recall | F1-score |
| Bugs (69) | 0.53 | 0.67 | 0.59 | 0.51 | 0.61 | 0.56 |
| Complex_data (102) | 0.51 | 0.80 | 0.62 | 0.64 | 0.63 | 0.63 |
| Flat_data (127) | 0.73 | 0.54 | 0.62 | 0.61 | 0.63 | 0.62 |
| Gable_hip_other (246) | 0.93 | 0.75 | 0.83 | 0.88 | 0.83 | 0.86 |
| Overall Accuracy (544) | — | — | 0.70 | — | — | 0.72 |
| Macro Average (544) | 0.67 | 0.69 | 0.67 | 0.66 | 0.67 | 0.67 |
| Weighted Average (544) | 0.75 | 0.70 | 0.71 | 0.73 | 0.72 | 0.72 |

Table 3. Classification report of ConvNeXt without and with callbacks

Table 3 presents the classification report of the ConvNeXt model. Without callbacks, the model achieved an overall accuracy of 70%, which improved to 72% when callbacks were applied. Performance on the Complex_data class notably improved with callbacks, increasing precision from 0.51 to 0.64.

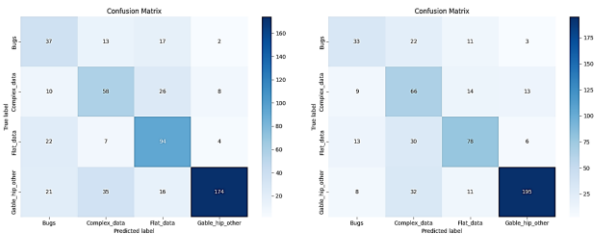
Table 4 presents the classification report of ResNet50 model on the test dataset. Without callbacks, the model attained an overall test accuracy of 53%, which slightly decreased to 51% when callbacks were applied. While the Gable_hip_other class remained the most accurately predicted in both cases having F1-score of 0.68 without callbacks and 0.70 with callbacks. Table 5 shows the test performance of the ResNet101 model under both training conditions. The overall test accuracy was 56% without callbacks and slightly declined to 52% when callbacks were applied. The model struggled with Bugs, especially without callbacks, where recall dropped to just 3%, and F1-score was only 0.05 indicating near-total failure in identifying that class.

| Class | Without callbacks | | | With callbacks | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1-score | Precision | Recall | F1-score |
| Bugs (69) | 0.42 | 0.38 | 0.40 | 0.00 | 0.00 | 0.00 |
| Complex_data (102) | 0.43 | 0.67 | 0.52 | 1.00 | 0.03 | 0.06 |
| Flat_data (127) | 0.36 | 0.34 | 0.35 | 0.35 | 0.46 | 0.40 |
| Gable_hip_other (246) | 0.74 | 0.62 | 0.68 | 0.58 | 0.88 | 0.70 |
| Overall Accuracy (544) | — | — | 0.53 | — | — | 0.51 |
| Macro Average (544) | 0.49 | 0.50 | 0.49 | 0.48 | 0.34 | 0.29 |
| Weighted Average (544) | 0.55 | 0.53 | 0.54 | 0.53 | 0.51 | 0.42 |

Table 4. Classification report of ResNet50 without and with callbacks

| Class | Without callbacks | | | With callbacks | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1-score | Precision | Recall | F1-score |
| Bugs (69) | 0.33 | 0.03 | 0.05 | 0.38 | 0.04 | 0.08 |
| Complex_data (102) | 0.56 | 0.42 | 0.48 | 0.52 | 0.48 | 0.50 |
| Flat_data (127) | 0.38 | 0.30 | 0.34 | 0.35 | 0.44 | 0.39 |
| Gable_hip_other (246) | 0.61 | 0.89 | 0.72 | 0.62 | 0.71 | 0.66 |
| Overall Accuracy (544) | — | — | 0.56 | — | — | 0.52 |
| Macro Average (544) | 0.47 | 0.41 | 0.40 | 0.47 | 0.42 | 0.41 |
| Weighted Average (544) | 0.51 | 0.56 | 0.50 | 0.51 | 0.52 | 0.49 |

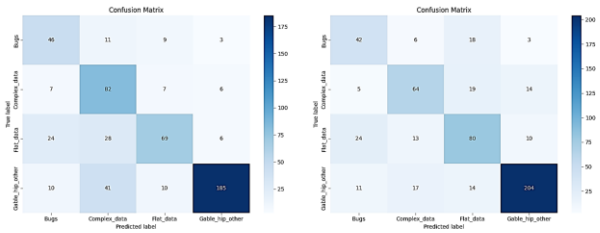
Table 5. Classification report of ResNet101 without and with callbacks




(a) (b)

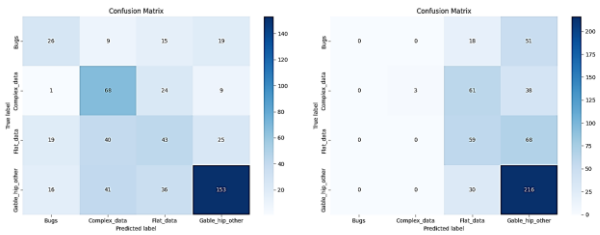
Figure 10. Confusion matrix of ViT Classifier (a) without and (b) with callbacks.

Figure 10 shows that the ViT Classifier with callbacks improved class-specific predictions with the help of confusion matrix. For instance, Gable_hip_other correctly classified 195 images with callbacks compared to 174 without. Similarly, Complex_data improved from 58 to 66 correct classifications. However, misclassifications in the Bugs class increased from 37 to 33, indicating persistent confusion in that category. Figure 11 shows that ConvNeXt improved the classification of Gable_hip_other from 185 to 204 correct predictions with callbacks. Flat_data also improved from 69 to 80 correct predictions, while Complex_data accuracy slightly dropped from 82 to 64 due to increased confusion with other classes. Figure 12 shows that ResNet50's performance on Bugs significantly dropped with callbacks, misclassifying all 69 images, while Gable_hip_other improved with 216 correct predictions. Flat_data also improved from 43 to 59 correct predictions, but overall class confusion increased in the callback case. Figure 13 shows that ResNet101 struggled with the Bugs class, correctly identifying only 2 to 3 images in both cases, while Gable_hip_other had the highest accuracy with 220 correct predictions without callbacks and 175 with callbacks, indicating performance degradation despite improved Flat_data recognition.



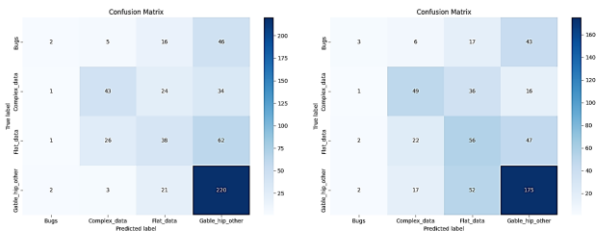

(a) (b)

Figure 11. Confusion matrix of ConvNeXt (a) without and (b) with callbacks.




(a) (b)

Figure 12. Confusion matrix of ResNet50 (a) without and (b) with callbacks.




(a) (b)

Figure 13. Confusion matrix of ResNet101 (a) without and (b) with callbacks.

### 3.4 Qualitative analysis

This section presents qualitative analysis of all models illustrating their performance in the case of callbacks and without them. In figures, green label shows the correct prediction of classes and red denotes the misclassification. Figure 14 shows that the ViT classifier demonstrates improved prediction consistency across most classes when callbacks are applied, though some misclassifications persist, especially in the

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-5-W2-2025-577-2025 | © Author(s) 2025. CC BY 4.0 License.

Bugs and Complex_data categories while there are issues in Flat_data class also when callbacks are applied.

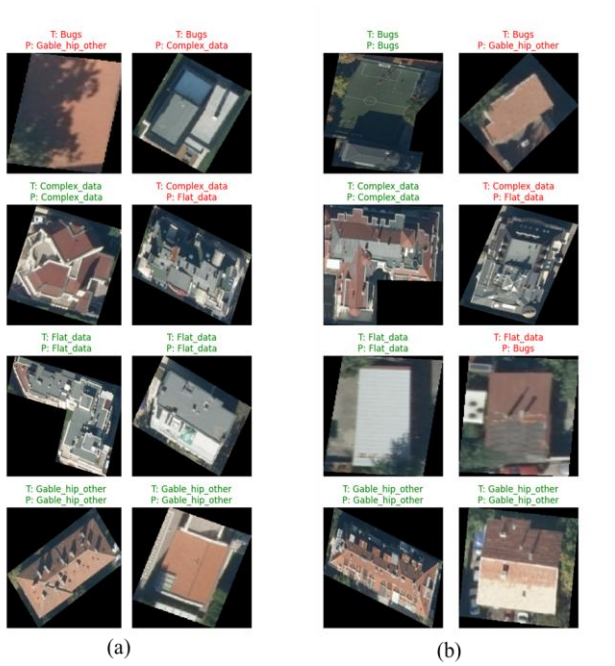



Figure 14. Sample prediction results of the ViT Classifier across all four classes with two test images per class (a) without and (b) with callbacks.

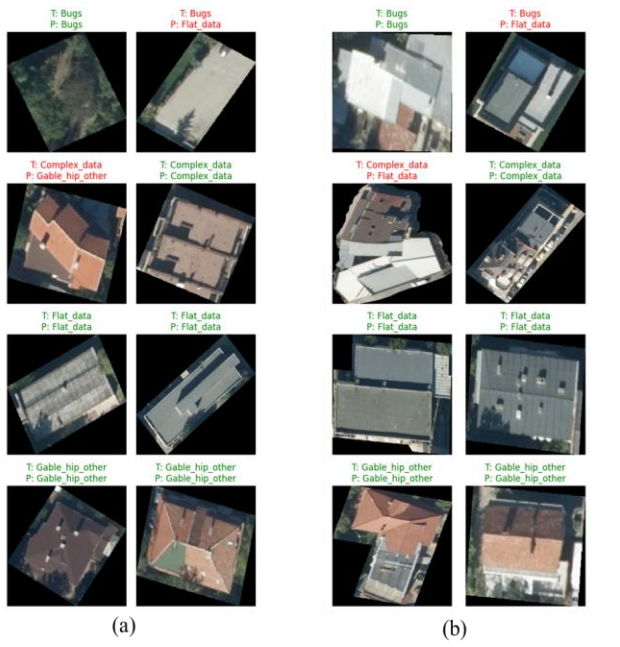



Figure 15. Sample prediction results of the ConvNeXt across all four classes with two test images per class (a) without and (b) with callbacks.

Figure 15 illustrates that ConvNeXt performs reliably across most classes, especially for Flat_data and Gable_hip_other, with minimal confusion. The use of callbacks still faced the issue of misclassification, notably for the Complex_data and Bugs classes. As shown in Figure 16, ResNet50 performed poorest in both the cases by misclassifying most of the classes but Gable_hip_other class when callbacks were used. Without the use of callbacks slightly improved the classification prediction for Complex_data results. Performance of ResNet101 is improved than ResNet50 with the use of callbacks but still faces issues in Bugs and Flat_data classes as shown in Figure 17. Overall, all the models performed well in case of predicting Gable_hip_other class and poorly predicted Bugs class.

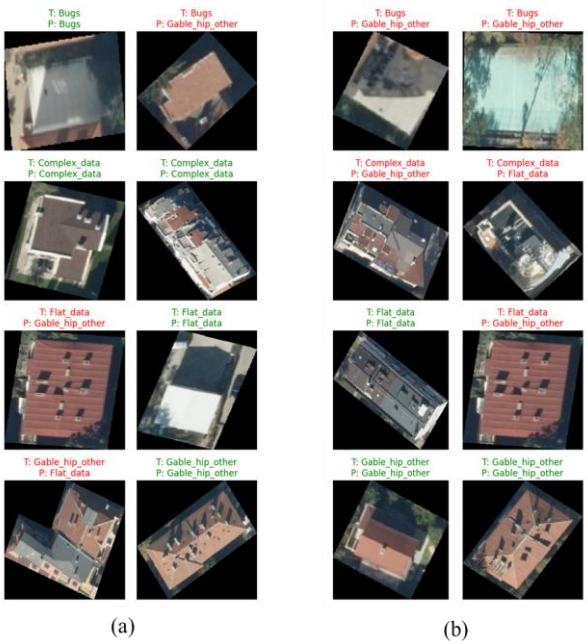



Figure 16. Sample prediction results of the ResNet50 across all four classes with two test images per class (a) without and (b) with callbacks.

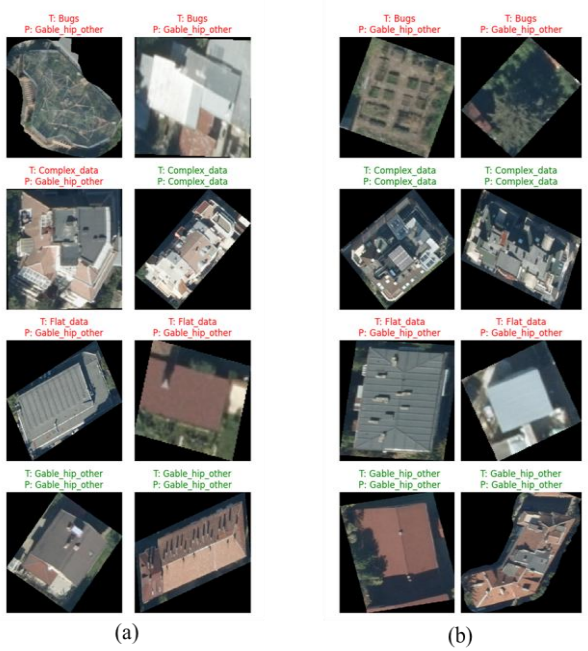



Figure 17. Sample prediction results of the ResNet101 across all four classes with two test images per class (a) without and (b) with callbacks.

## 4. Conclusion

This study presented an effective transformer based deep learning model for rooftop type classification from orthophotos,

describing the capabilities of the Swin Transformer architecture. By exploiting its hierarchical feature representation and self-attention mechanism, the Swin Transformer demonstrated superior performance in accurately distinguishing between four rooftop classes—Bugs, Complex_data, Flat_data, and Gable_hip_other. Extensive experiments were conducted to evaluate the effectiveness of callbacks during training, which contributed to improved generalization and avoid overfitting. A comprehensive comparison with other prominent convolutional and transformer-based models including ViT Classifier, ConvNeXt, ResNet50 and ResNet101 revealed that the Swin Transformer consistently outperformed across various evaluation metrics. Both quantitative results such as classification reports and confusion matrices and qualitative visualizations confirmed that the Swin Transformer could effectively handle variations in rooftop shapes and textures. Future research involves varied rooftop types on Indian datasets and refining model architectures to address challenges posed by similar-looking or spatially overlapping rooftops.

## References

Alidoost, F., & Arefi, H. (2019). A CNN - Based Approach for Automatic Building Detection and Recognition of Roof Types Using a Single Aerial Image. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, *86*(5), 235–248. https://doi.org/10.1007/s41064-018-0060-5

Biljecki, F., Stoter, J., Ledoux, H., Zlatanova, S., & Çöltekin, A. (2015). Applications of 3D City Models : State of the Art Review. *ISPRS International Journal of Geo-Information*, 2842–2889. https://doi.org/10.3390/ijgi4042842

Buyukdemircioglu, M., Can, R., & Kocaman, S. (2021). DEEP LEARNING BASED ROOF TYPE CLASSIFICATION USING VERY HIGH RESOLUTION AERIAL IMAGERY. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLIII-B3-2*, 55–60. https://doi.org/10.5194/isprs-archives-XLIII-B3-2021-55-2021

Castagno, J., & Atkins, E. (2018). Roof Shape Classification from LiDAR and Satellite Image Data Fusion Using Supervised Learning. *Sensors*, *18*(11), 3960. https://doi.org/10.3390/s18113960

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE. *ICLR*.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 770–778. https://doi.org/10.48550

Hristov, E., Petrova-Antonova, D., Petrov, A., Borukova, M., & Shirinyan, E. (2023). *Imagery dataset for rooftop detection and classification*. https://doi.org/10.5281/zenodo.7633594

Li, L., Song, N., Sun, F., Liu, X., Wang, R., Yao, J., & Cao, S. (2022). Point2Roof : End-to-end 3D building roof modeling from airborne LiDAR point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, *193*, 17–28. https://doi.org/10.1016/j.isprsjprs.2022.08.027

Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). *Thresholding Classifiers to Maximize F1 Score*. http://arxiv.org/abs/1402.1892

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. https://arxiv.org/abs/2103.14030

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR52688.2022.01167

Mohajeri, N., Assouline, D., Guiboud, B., & Bill, A. (2018). A city-scale roof shape classi fi cation using machine learning for solar energy applications. *Renewable Energy*, *121*, 81–93. https://doi.org/10.1016/j.renene.2017.12.096

O'Shea, K., & Nash, R. (2015). An Introduction to Convolutional Neural Networks. *International Journal for Research in Applied Science and Engineering Technology*, *10*(12), 943–947. https://doi.org/10.48550

Partovi, T., Fraundorfer, F., Marmanis, D., & Reinartz, P. (2017). ROOF TYPE SELECTION BASED ON PATCH-BASED CLASSIFICATION USING DEEP LEARNING FOR HIGH RESOLUTION SATELLITE IMAGERY. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLII-1/W1*, 653--657. https://doi.org/10.5194/isprs-archives-XLII-1-W1-653-2017

Prechelt, L. (1998). Automatic early stopping using cross validation: quantifying the criteria. Neural Networks, 11(4), 761–767.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning representations by back-propagating errors. Nature, 323, 533-536.

Smith, L. N. (2017). Cyclical learning rates for training neural networks. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 464–472. https://doi.org/10.1109/WACV.2017.58

Story, M., & Congalton, R. G. (1986). Accuracy Assessment: A User's Perspective. *Photogrammetric Engineering and Remote Sensing*, *52*(3), 397–399. https://www.asprs.org/wp-content/uploads/pers/1986journal/mar/1986_mar_397-399.pdf

Vaswani, A. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*.

Wang, Y., Li, S., Teng, F., Lin, Y., Wang, M., & Cai, H. (2022). Improved Mask R-CNN for Rural Building Roof Type Recognition from UAV High-Resolution Images : A Case Study in Hunan Province , China. *Remote Sensing*, *14*(2), 265. https://doi.org/10.3390/rs14020265

Yildirim, M., & Karsli, F. (2024). A Novel Deep Learning Based Classification of Building Roof Types Using Point A Novel Deep Learning Based Classification of Building Roof Types Using Point Cloud Data. *Journal of the Indian Society of Remote Sensing*, *October*. https://doi.org/10.1007/s12524-024-01986-z