

Deep Learning Architectures for Building Extraction from UAV Imagery: A Comparative Analysis of Transfer Learning and Data Augmentation Strategies

TRUNG Dung Pham^{1,*}, NGOC Hung Pham²

¹ Hanoi University of Mining and Geology, Hanoi, Vietnam

² Phenikaa University, Hanoi, Vietnam

* Corresponding author: phamtrungdung@hmg.edu.vn

Abstract: Automated building extraction using deep learning's semantic segmentation techniques on remote sensing imagery is increasingly important for mapping, urban planning, and environmental monitoring. However, this task presents considerable challenges, particularly when using Unmanned Aerial Vehicle (UAV) imagery from developing countries like Vietnam. The construction landscape in Vietnam is often characterized by non-homogeneous and non-standard building designs, as well as the coexistence of slums and modern urban areas. These real-world issues, combined with a lack of annotated data, frequently lead to model overfitting and reduced generalization. This study aims to fill a critical knowledge gap by systematically evaluating six widely used deep learning architectures: DeepLabV3+, Unet, Unet++, Feature Pyramid Network (FPN), Pyramid Attention Network (PAN), and SegFormer, specifically for building extraction tasks. We utilize Intersection over Union (IoU) scores and analyze overfitting behavior to compare the performance of these models on a benchmark WHU Building Dataset and a challenging proprietary Vietnamese UAV Building Dataset. Our thorough analysis quantifies the individual and combined effects of transfer learning and data augmentation on model performance. Our findings indicate that both transfer learning and data augmentation are essential, consistently resulting in higher IoU scores improving by 10-15% with transfer learning—and significantly reducing overfitting (from as much as 30% down to 3-7% on the Vietnamese dataset). Furthermore, architectural efficiency is more critical than the raw parameter count. Models like DeepLabV3+ and SegFormer achieved top-tier performance, with IoU scores of 0.858 and 0.855, respectively, on the Vietnamese dataset while maintaining a moderate parameter count of 11-12 million. The Vietnamese dataset proved to be inherently more challenging than the WHU benchmark, highlighting the need for robust training strategies for effective generalization in complex, real-world environments. This research provides recommendations for selecting appropriate model architectures and training strategies to ensure accurate and reliable building extraction under difficult conditions.

Keywords: Building Extraction, Deep Learning, UAV Imagery, Transfer Learning, Data Augmentation

1. Introduction

Automated building extraction for the diversity of building structures using ultra-high resolution UAV images is a critical task with wide-ranging applications in urban planning (Freire et al., 2014), environmental monitoring (Sahar et al., 2010), and cartography (Lafarge et al., 2008). Deep learning, particularly semantic segmentation, has emerged as the leading approach for this challenge, demonstrating superior capabilities in pixel-wise classification (Liu et al., 2022). However, this task presents unique complexities in Vietnam, where building structures are often non-homogeneous, non-standard in construction, and slums frequently coexist with modern urban areas. The performance of these models is significantly influenced by architectural choices (Mayer, 1999), training methodologies (Luo et al., 2021), and the inherent complexity of the datasets used for training and evaluation. Real-world remote sensing data, such as Unmanned Aerial Vehicle (UAV) imagery, often presents unique challenges, including high architectural diversity, occlusions from dense vegetation, varied lighting conditions, and limited annotated

samples (Luo et al., 2023), which can lead to issues like overfitting (Chen & Lu, 2019) and reduced generalization capabilities.

Despite the advancements in deep learning for semantic segmentation, a knowledge gap persists regarding the systematic evaluation of diverse model architectures and the quantitative impact of specific training strategies, such as transfer learning and data augmentation (Fayaz et al., 2024; Marastoni et al., 2021; Zhang & Tang, 2018) on both segmentation accuracy (IoU) and overfitting behavior, especially when transitioning from controlled benchmark datasets to complex real-world scenarios. This gap creates uncertainty in selecting optimal models and training paradigms for practical building extraction applications.

To address this, our study comprehensively investigates the performance of several advanced deep learning architectures for semantic image segmentation in the context of building extraction. Our primary objective is to analyze both Intersection over Union (IoU) scores and overfitting behavior across a diverse set of models, including DeepLabV3+ (Chen et al., 2018), Unet (Ronneberger et al., 2015), Unet++ (Zhou et al., 2018), Feature Pyramid Network (FPN) (Lin et al., 2017), Pyramid Attention Network (PAN) (Li et al., 2018), and SegFormer (Xie et al., 2021). We evaluate these models on two distinct building datasets: a subset of the widely recognized WHU Building Dataset (Ji et al., 2018), serving as a benchmark for comparative analysis, and a challenging proprietary Vietnamese UAV Building Dataset, designed to assess real-world applicability under complex conditions. A core focus of our research is to conduct an ablation analysis of key training strategies, specifically transfer learning and data augmentation, to quantify their individual and synergistic contributions to model accuracy and robustness.

This research aims to answer several key questions: How do we evaluate the difficulty of a real-world dataset compared to a template dataset (e.g., WHU building dataset)? How large is the improvement in segmentation accuracy in a real-world dataset achieved by the training strategies of data augmentation and transfer learning? How important are model complexity and architectural design to the performance of segmentation tasks in complex and limited datasets? Ultimately, this study seeks to identify the optimal strategies for selecting model architectures and training methodologies to achieve robust and generalizable building extraction in diverse, challenging real-world environments.

Our findings reveal that while architectural design is crucial, the choice of training methodology profoundly impacts performance. We demonstrate that the combined application of transfer learning and data augmentation is indispensable, consistently yielding superior IoU scores and significantly mitigating overfitting across all models and both datasets. Notably, models like DeepLabV3+ and SegFormer exhibit high architectural efficiency, achieving top-tier performance with moderate parameter counts, making them suitable for practical deployment. Furthermore, our comparative analysis explicitly quantifies the increased difficulty of real-world UAV datasets compared to benchmark equivalents, underscoring the necessity of robust training strategies for effective generalization.

The remainder of this paper is structured as follows: Section 2 details the datasets, model architectures, and experimental setup. Section 3 presents the comprehensive results, including performance on the WHU dataset, an ablation analysis on the Vietnamese UAV dataset, and a comparative assessment of dataset difficulty. Section 4 provides a discussion of our findings, interpreting the implications for model selection and training strategies. Finally, Section 5 concludes the study, summarizing key contributions, practical applications, limitations, and avenues for future research.

2. Materials and Methods

2.1. Dataset

Our study employed two distinct building extraction datasets to ensure both benchmark comparability and robust real-world validation.

The WHU Building Dataset (Ji et al., 2018) was utilized for international benchmarking and comparison with existing state-of-the-art methods. This widely recognized dataset features high-resolution

aerial imagery over Christchurch, New Zealand, at 0.3 meters per pixel, serving as a standard in remote sensing segmentation. For our comparative study, a 60% subset of the original WHU dataset (2843 training samples, 1053 validation samples) was carefully selected. This subset maintained representative diversity while efficiently managing computational resources across multiple model comparisons. The WHU dataset is characterized by its relatively homogeneous architectural and environmental context.

Conversely, the Vietnamese UAV Building Dataset was used to evaluate model performance under challenging, real-world conditions. This proprietary dataset comprises ultra-high-resolution imagery (0.1 meters per pixel) from various provinces across northern, central, and southern Vietnam. It was intentionally designed as an exceptionally difficult testbed due to its unique and dynamic characteristics. The dataset captures significant architectural diversity, including industrial areas, modern high-rise, and traditional rural dwellings. Images are further complicated by dense vegetation occlusions, overlapping structures, high spectral variability from diverse building materials, and inconsistencies from varied UAV campaigns and weather conditions. This starkly contrasts with the WHU dataset's more homogeneous urban environment, making the Vietnamese dataset critical for assessing model robustness in truly complex, real-world scenarios.

In essence, the WHU subset provided a controlled environment for initial benchmarking and resource management, while the Vietnamese dataset offered rigorous, real-world validation to demonstrate the practical robustness of the models.

2.2. Model Architectures

The six model architectures (U-Net, U-Net++, DeepLabV3+, FPN, PAN, and SegFormer) were chosen for this comprehensive evaluation. A short description of these model architectures is as follows:

Unet (Ronneberger et al., 2015): A foundational architecture in biomedical image segmentation, Unet employs a symmetric encoder-decoder structure. The main idea behind this architecture is the use of direct skip connections that concatenate features from the encoder to corresponding decoder stages. This mitigates spatial information loss during downsampling, crucial for precise boundary extraction.

Unet++ (Zhou et al., 2018): An evolution of Unet, Unet++ introduces nested and dense skip pathways between the encoder and decoder. This provides a more flexible and robust feature fusion mechanism. This architecture enables richer multi-scale feature aggregation and potentially more accurate segmentation.

DeepLabV3+ (Chen et al., 2018): This architecture combines the strengths of encoder-decoder models with atrous convolutions (dilated convolutions). Atrous Spatial Pyramid Pooling (ASPP) is used to extract contextual information at the encoder stage. ASPP can capture multi-scale features by applying atrous convolutions at different rates. The decoder then reconstructs fine-grained spatial details, enabling precise segmentation.

Feature Pyramid Network (FPN) (Lin et al., 2017): FPN is designed to handle objects across a wide range of scales by constructing a feature pyramid. It employs a top-down pathway to upsample high-level semantic features from deep encoder layers, combined with lateral connections that merge these with fine-grained details from shallow layers. This creates a rich, multi-scale feature representation suitable for segmenting objects of widely varying sizes.

Pyramid Attention Network (PAN) (Li et al., 2018): Building upon the idea of feature pyramids, PAN enhances feature fusion by integrating attention mechanisms. It typically combines a Feature Pyramid Attention (FPA) module in the encoder (to capture global contextual information) with a Pixel Attention (PA) module in the decoder (to refine pixel-wise predictions), aiming to accurately delineate the boundaries of objects and filter out irrelevant noise in segmentation results.

SegFormer (Xie et al., 2021): A simple and efficient segmentation architecture, SegFormer combines a hierarchical Transformer encoder and a simple Multi-Layer Perceptron (MLP) decoder. The hierarchical Transformer encoder captures global contextual information at multiple scales using self-attention, while

the simple MLP decoder combines information from different layers of the encoder, which helps in creating powerful representations for accurate segmentation.

2.3. Transfer Learning with Pre-trained Weights

Transfer learning is a powerful technique in deep learning that involves leveraging knowledge gained from solving one problem and applying it to a different but related problem. In our study, this is achieved by using a model's encoder (backbone) that has been pre-trained on a large-scale, general-purpose image dataset, specifically ImageNet (Deng et al., 2009). Pre-training on ImageNet allows the encoder to learn a rich hierarchy of both low-level and high-level visual features, such as edges, textures, and shapes. These generalized features are highly effective for object recognition and are then used as a starting point for the specific task of building segmentation. Transfer learning significantly improves initial feature extraction capabilities, accelerates model convergence during training, and enhances performance, particularly when working with domain-specific datasets. The Vietnamese UAV building dataset is an example of limited samples or high diversity of landscape.

2.4. Data Augmentation

To enhance model robustness and generalization, an extensive data augmentation pipeline was implemented during training. Augmentation (Wang et al., 2022) is particularly necessary for a dataset like the Vietnamese UAV building dataset. Since its sample size is limited relative to the vast number of parameters in deep learning models, as well as the diversity in building structures and landscapes. Without these techniques, models are prone to overfitting by simply memorizing the specific input images, which leads to poor performance on unseen real-world data.

In our study, the geometric transformations (Li et al., 2024) are applied, which are a highly suitable method for augmenting RGB images in building extraction tasks. This transformation can preserve the essential shapes and structures of buildings while introducing realistic variations. This primarily focused on geometric transformations, including random horizontal and vertical flipping, as well as random rotations (see Tab. 1). These techniques effectively increased the diversity of the training data and reduced the risk of overfitting to specific spatial patterns or building orientations.

Tab. 1. The parameter for the Geometric transformation in the data augmentation technique

Augmentation Technique	Parameter / Limit	Probability (p)	Notes
Resize	width=trainsize, height=trainsize	N/A	Resizes images to a consistent input dimension (e.g., 512x512).
ShiftScaleRotate	Shift: 0.05, Scale: 0.1, Rotate: 15	0.8	Randomly shifts (up to 5% of dimensions), scales ($\pm 10\%$), and rotates ($\pm 15^\circ$).
HorizontalFlip	N/A	0.5	Randomly flips images horizontally.
VerticalFlip	N/A	0.3	Randomly flips images vertically, useful for orthogonal views.

2.5. Experimental Setup

2.5.1. Computer Configuration

All experiments were conducted using Google Colaboratory's free-tier computational resources. Google Colaboratory provides access to virtual machines equipped with Graphics Processing Units (GPUs) for accelerated deep learning computations. The specific hardware configuration allocated to users in the free-tier can vary dynamically based on availability and usage patterns.

During the experimental period, the GPU provided was an NVIDIA Tesla T4 with 15360 MiB (15 GB) of GDDR6 Video RAM (VRAM). The NVIDIA driver version was 550.54.15, and the CUDA version

was 12.4. The virtual machine instances are generally powered by an Intel Xeon CPU operating at approximately 2.20 GHz, configured with 2 virtual CPUs (vCPUs). System memory (RAM) is typically around 13 GB. For storage, a temporary disk space of approximately 60-100 GB is provided, which is reset after each session. The environment runs on a Linux-based operating system and comes with popular deep learning frameworks such as TensorFlow and PyTorch pre-installed.

2.5.2. Experimental Setup

The hyperparameters used for model training were carefully selected to achieve a balance between precision, stability, and computational efficiency. These parameters, as detailed in Tab. 2, form a robust training regime for building extraction. A maximum of 50 epochs was established to ensure models had sufficient time to converge without excessive computational cost. A batch size of 8 was chosen to balance memory constraints with the need for stable gradient estimates, which is particularly important for generalization on diverse building shapes. The AdamW optimizer was utilized for its proven effectiveness in handling sparse gradients and providing adaptive learning rates. Furthermore, a StepLR scheduler was implemented to aggressively decay the learning rate, which is a common practice for refining boundaries and improving fine-grained segmentation after the initial training phase. Finally, an early stopping mechanism with a patience of 30 epochs and a min_delta of 0.001 was configured to prevent premature stopping and facilitate the model's gradual refinement toward optimal accuracy. These parameters are used to conduct a fair comparison but not to optimize the model.

Tab. 2. Hyperparameters for training the Model

Parameter	Value	Notes
Maximum Epochs	50	Maximum number of training iterations.
Batch Size	8	Number of images processed per training step.
Initial Learning Rate	1e-4	Starting learning rate for the optimizer.
Optimizer	AdamW	Adaptive moment estimation with weight decay regularization.
Learning Rate Scheduler	StepLR	Step decay with step_size=30 epochs and gamma=0.1.
Loss Function	BCEWithLogitsLoss	Binary Cross-Entropy with logits for binary segmentation.
Early Stopping Patience	30	Number of epochs to wait for improvement before stopping.
Early Stopping Min Delta	0.001	Minimum change in validation IoU to qualify as improvement.

2.5.3. Evaluation Metric

Model performance was quantitatively evaluated using Intersection over Union (IoU) (Yang et al., 2018). IoU is a standard metric for semantic segmentation, measuring the spatial overlap between the predicted building mask and the ground truth mask. It is calculated as the area of intersection divided by the area of the union of the predicted and ground truth segmentation regions. A higher IoU value indicates better segmentation accuracy and boundary delineation. The formula for IoU is given by:

$$\text{IoU} = \text{TP} / (\text{TP} + \text{FP} + \text{FN}) \quad (1)$$

where True Positives (TP) represent pixels correctly identified as buildings, False Positives (FP) occur when non-building pixels are erroneously identified as buildings, and False Negatives (FN) represent building pixels incorrectly classified as non-building areas.

IoU (Intersection over Union) is the most effective metric for semantic segmentation because it accurately measures the spatial overlap between predicted and ground truth masks—the core objective of the task. Other metrics like accuracy, precision, recall, and F1-score can be misleading and fail to capture geometric correctness, especially in cases of class imbalance. For instance, accuracy may appear high even if no objects are detected in an imbalanced dataset. Meanwhile, precision, recall, and F1-score do not

penalize poorly shaped masks. This means a model could achieve high scores despite imprecise predictions. In this context, IoU stands out as essential for ensuring the quality and reliability of semantic segmentation outcomes. Besides, due to the limited scope of the presented text, other metrics were not included in this study.

3. Results

This section presents the experimental results obtained from evaluating six deep learning models—DeepLabV3+, U-Net, U-Net++, FPN, PAN, and SegFormer—on both the subset of the WHU Building Dataset and the Vietnamese UAV Building Dataset. Performance was primarily assessed using the Intersection over Union (IoU) metric, alongside an analysis of overfitting behavior. Noting that the IoU value is averaged over three times during the training process. This analysis aims to highlight the effectiveness of different model architectures and training strategies, particularly when applied to a challenging real-world building dataset with diverse constructions and limited samples.

3.1. Performance on the WHU Building Dataset (60% Subset)

Evaluating the models on the subset of the WHU Building Dataset establishes a benchmark for their performance under relatively controlled conditions (see Fig. 1 and Fig. 2). IoU scores for models trained solely from scratch (Without DA) were notably lower, with PSPNet exhibiting the lowest at 0.761 IoU. The consistent application of Data Augmentation (Scratch with DA) across all models improved IoU, demonstrating its effectiveness in enhancing generalization. However, the most substantial gains were observed with the combined application of Data Augmentation and Transfer Learning (Transfer learning with DA).

In this fully optimized "Transfer learning with DA" configuration, Unet++ achieved the highest IoU of 0.891, showcasing its strong adaptability and superior performance in complex imagery when paired with these training strategies. DeepLabV3+ and FPN also performed very well, with IoU scores of 0.877 and 0.880, respectively. SegFormer recorded a competitive IoU of 0.871 in this setup. While PSPNet improved with "Transfer learning with DA" (0.868 IoU), its performance remained comparatively lower in this advanced configuration.

Transfer Learning methods (both with and without Data Augmentation) consistently and significantly outperformed training from scratch, yielding IoU scores generally 10-15% higher across all model architectures. This highlights the substantial value of leveraging pre-trained weights for semantic segmentation tasks. Data Augmentation (DA) provided modest yet consistent improvements across all training methods, typically adding 1-3% to IoU scores, indicating its role in enhancing generalization.

Model Architecture Analysis: DeepLabV3 emerged as a top performer across all training scenarios on the WHU dataset, consistently achieving high IoU scores (ranging from 0.844 to 0.877) while maintaining relatively moderate parameter counts. Transformer-based models like SegFormer also demonstrated strong performance, particularly in transfer learning scenarios, reaching competitive IoU scores (0.878-0.879). Conversely, Unet consistently underperformed relative to its parameter count, suggesting it may not be the optimal choice for this dataset, especially when trained from scratch.

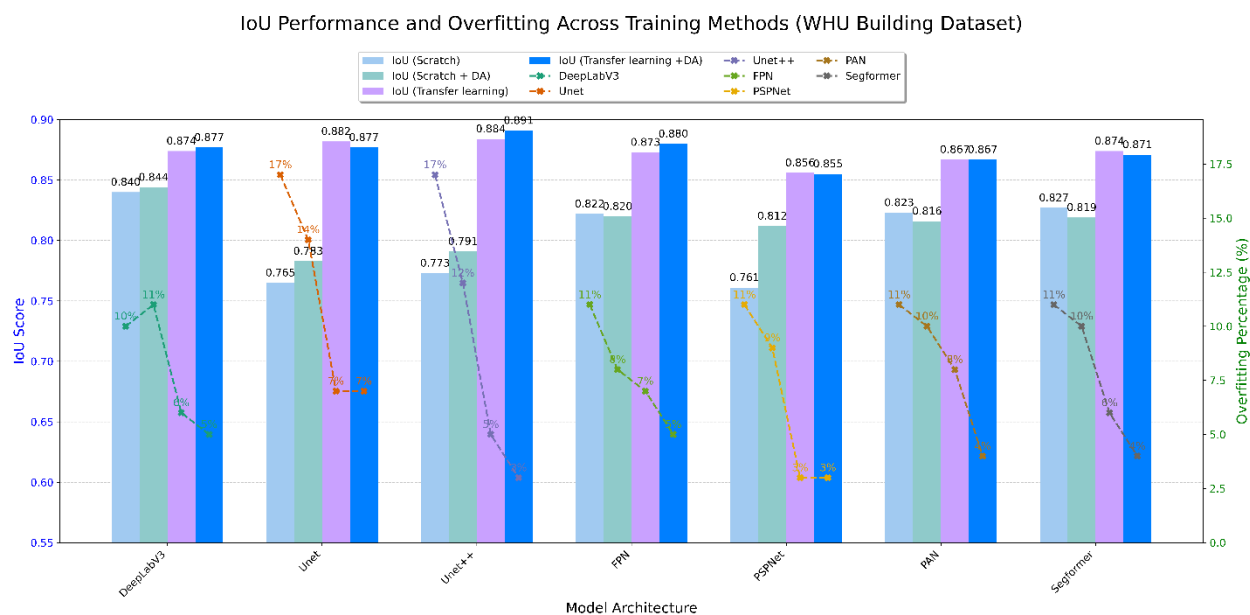


Fig. 1. IoU performance and overfitting across model architectures in the subset WHU building dataset

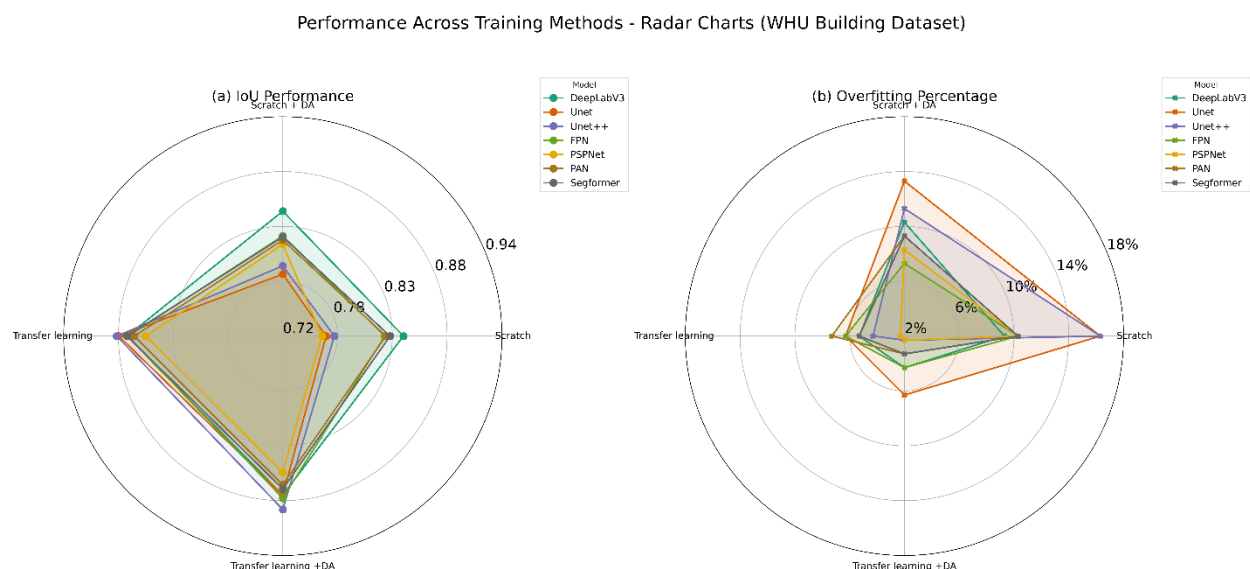


Fig. 2. Comparison of both IoU and overfitting values of the model architecture in four cases of training strategies in the subset WHU building dataset

Parameter Efficiency Insights: Our analysis revealed no clear direct correlation between model complexity (as indicated by parameter count) and IoU performance. Instead, architectural efficiency emerged as a pivotal determinant. Some models achieved excellent results with fewer parameters, while others required more for similar performance, emphasizing that the inherent design and efficiency of an architecture are often more critical than its raw size. Notably, models such as DeepLabV3+ and SegFormer achieved the highest IoU scores (0.858 and 0.855, respectively) while possessing a comparatively moderate number of parameters (ranging from 11 to 12 million).

Training Method Recommendations: For this WHU dataset, the combination of Transfer Learning and Data Augmentation appears to be the optimal approach. Within this configuration, Unet++, DeepLabV3+, and SegFormer stand out as the top architectural choices, depending on whether the primary focus is on absolute performance or a balance with parameter efficiency.



Fig. 3. The relationship between the parameters of model architecture and their performance across training methods

3.2. Ablation Analysis of Training Strategies on Vietnamese UAV Building Dataset

This section presents a comprehensive ablation analysis, examining the individual and combined contributions of Data Augmentation (DA) and Transfer Learning (TL) to both model performance (IoU) and overfitting behavior on the challenging Vietnamese UAV Building Dataset. This provides critical insights into effective strategies for real-world applications with limited and diverse data (see Fig. 4 and Fig. 5).

Baseline Performance and Overfitting: Training from Scratch (Without DA): When models were trained purely from scratch without any data augmentation or pre-trained weights, their performance on the complex Vietnamese UAV Building Dataset was generally the lowest. IoU scores in this baseline configuration ranged from 0.650 (PSPNet and PAN) to 0.740 (DeepLabV3+). DeepLabV3+ achieved the highest IoU in this scenario, suggesting some inherent robustness in its architecture even from random initialization, considering its relatively higher number of parameters (12.3M). Correspondingly, models trained purely from scratch exhibited the highest overfitting percentages, ranging from 21% (DeepLabV3+, Unet++, FPN) to a significant 30% (Unet, PAN, Segformer). This pronounced overfitting is a direct consequence of training complex deep learning models with a limited sample size compared to the vast number of parameters in these networks (ranging from 11.3M for PSPNet to 15.9M for Unet++).

Impact of Data Augmentation (DA) (Scratch + DA vs. Scratch): The introduction of Data Augmentation (DA) significantly improved IoU performance across all models when compared to their "Scratch (Without DA)" counterparts. IoU scores for models trained with DA (while still from scratch in terms of pre-trained weights) ranged from 0.736 (Segformer) to 0.794 (DeepLabV3+). The overfitting percentages in the "Scratch + DA" configuration were consistently and substantially lower than those from "Scratch (Without DA)" training, dropping to a range of 8% (FPN) to 14% (Unet). This significant reduction (e.g., Unet's overfitting dropped from 30% to 14%, Segformer from 30% to 10%) demonstrates that by artificially increasing the dataset's apparent size and variability, DA prevents the model from merely memorizing specific input instances.

Impact of Transfer Learning (TL) (Transfer learning without DA vs. Scratch without DA): The analysis of Transfer Learning (TL) without Data Augmentation reveals its direct contribution to model performance and overfitting mitigation even in isolation. IoU scores for models trained with TL but without DA ranged from 0.769 (Unet) to 0.836 (DeepLabV3+). Regarding overfitting, applying Transfer Learning alone (without DA) also contributed to reducing the issue, with percentages ranging from 12% (U-Net++) to 17% (Unet, PAN). While generally lower than "Scratch (Without DA)", it's noteworthy that for some models (e.g., FPN), the overfitting percentage with "TL without DA" (15%) might be slightly higher than with "Scratch + DA" (8%).

Effect of Data Augmentation and Transfer Learning (DA+TL): The most substantial improvements in both IoU performance and overfitting mitigation were consistently observed when both Data Augmentation and Transfer Learning (DA+TL) were combined. In this optimal configuration, IoU scores reached their highest, ranging from 0.814 (Unet) to a leading 0.858 (DeepLabV3+). This profound leap in IoU from all other configurations underscores the synergistic power of these two strategies for achieving peak performance in complex real-world scenarios. This combined approach also resulted in the lowest overfitting percentages across all models, ranging from a remarkable 3% (Unet++, PSPNet) to 7% (Unet).

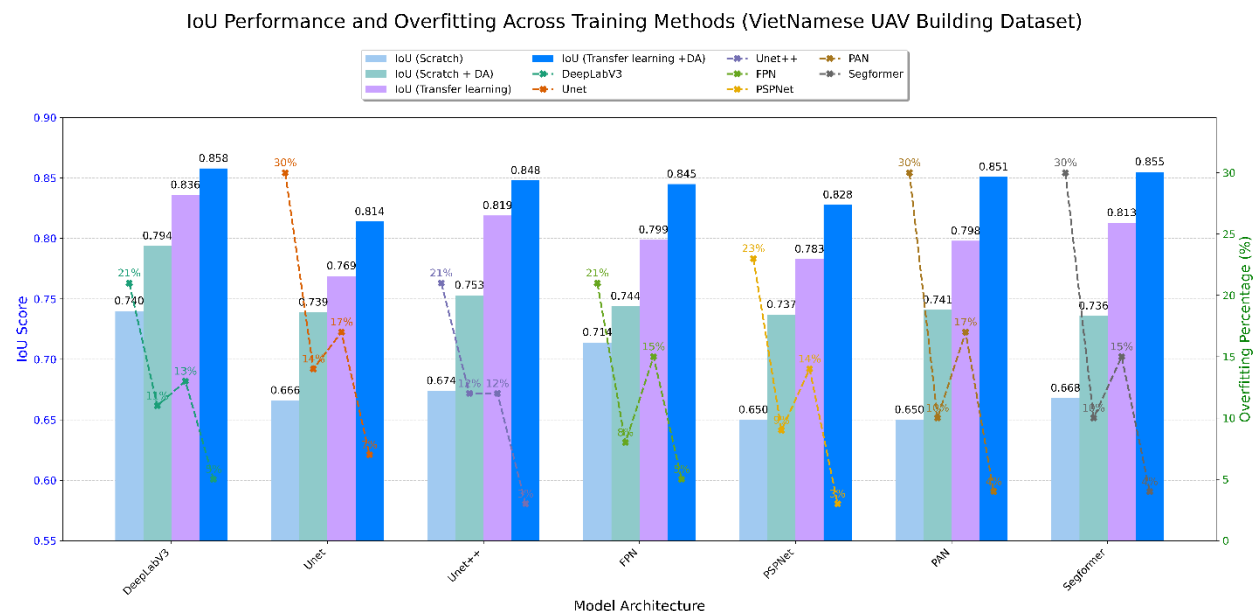


Fig. 4. IoU performance and corresponding overfitting across model architectures in the Vietnam UAV building dataset

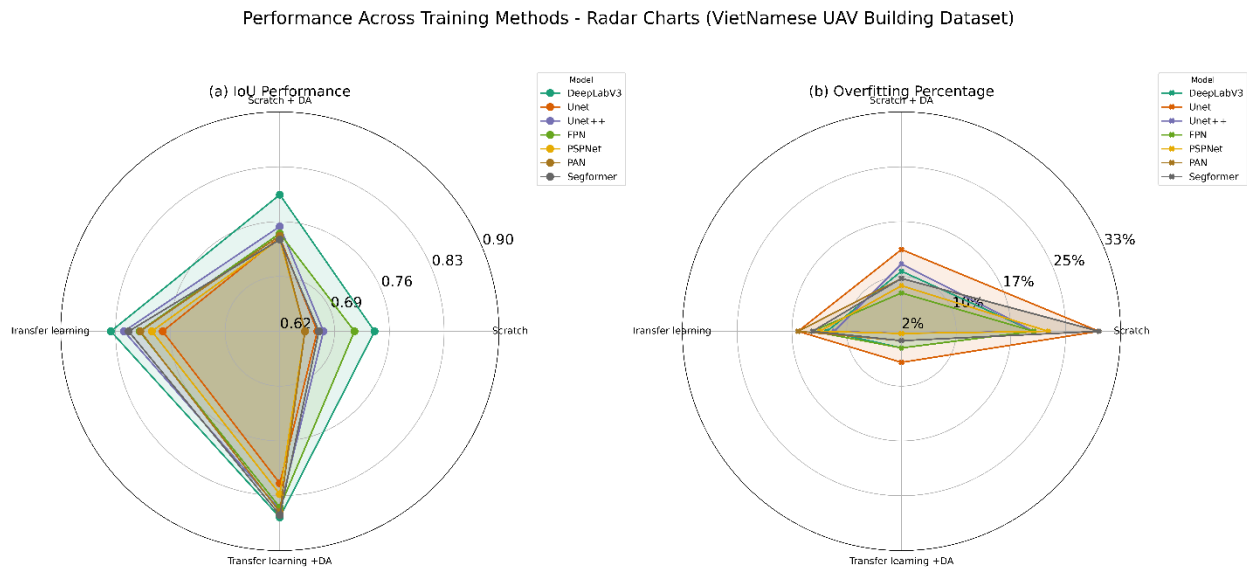


Fig. 5. Comparison of both IoU and overfitting values of the model architecture in four cases of training strategies in the Vietnamese UAV Building Dataset

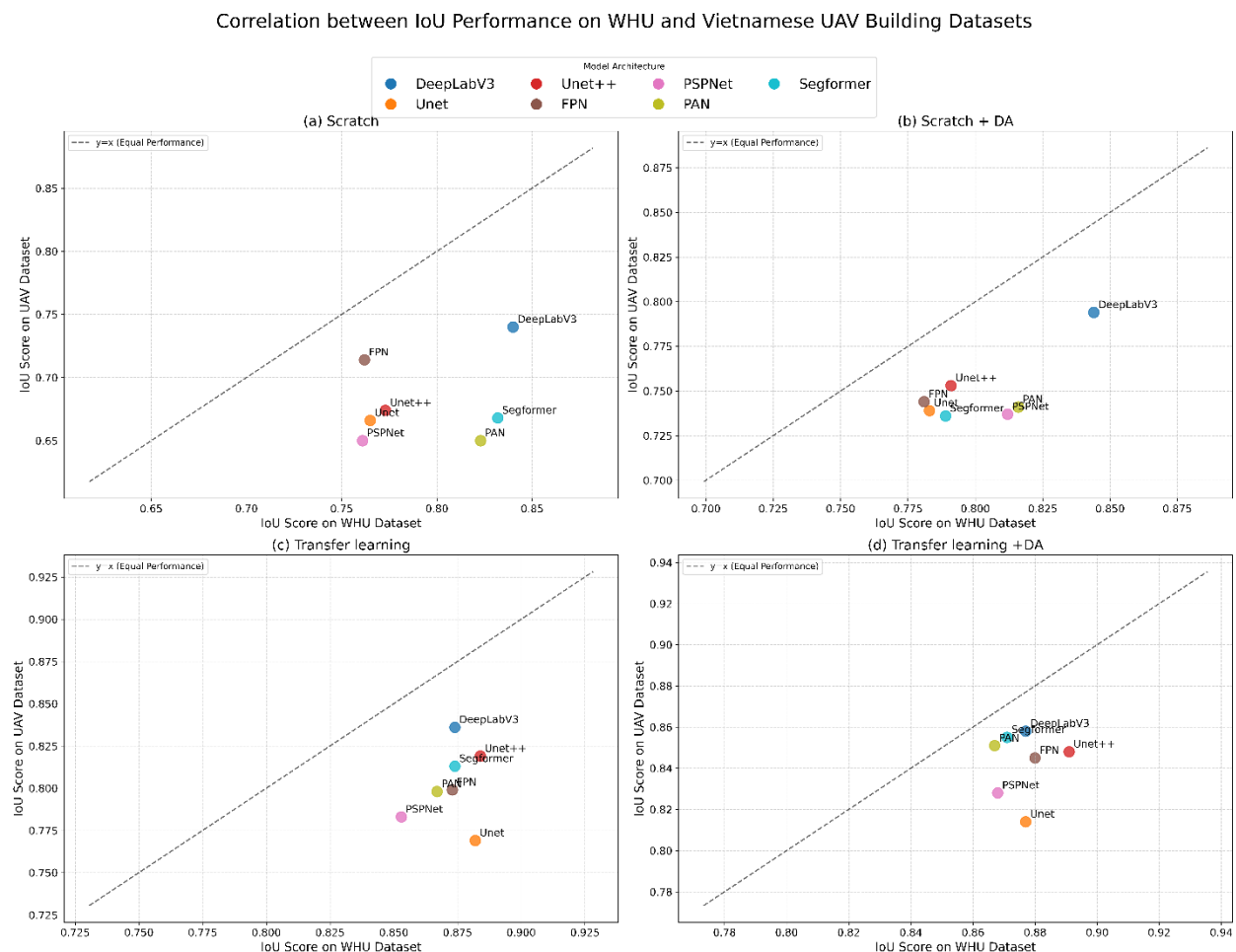
3.3. Comparative Analysis of Dataset Difficulty

The correlation plots (Fig. 6), which meticulously compare IoU scores for each model across the WHU and Vietnamese UAV Building Datasets under various training conditions, provide compelling evidence for the inherent difficulty of the real-world dataset. A primary and consistent observation is that, for every training condition evaluated, all data points are positioned below the diagonal line ($y=x$). This phenomenon consistently indicates that every model, regardless of its specific architecture or training methodology, achieves a lower IoU score on the Vietnamese dataset compared to the WHU dataset. This persistent performance gap serves as the most direct and quantifiable evidence that the Vietnamese dataset is fundamentally more challenging. The increased difficulty can be systematically attributed to specific real-world complexities that are less prevalent in the more homogeneous WHU dataset, such as high architectural diversity, frequent occlusions from vegetation, and the environmental variability resulting from image collection across multiple campaigns under different weather conditions. While this gap underscores the difficulty, the graphs also highlight the effectiveness of advanced training strategies. The application of Transfer Learning and Data Augmentation consistently pushes all models to achieve their highest IoU scores on both datasets, thereby demonstrating that a robust methodological approach is the most effective way to help models generalize and perform reliably, even when confronted with the inherent complexities of a real-world environment.

3.4. Qualitative Analysis and Visual Comparison of Training Strategies

While quantitative metrics like Intersection over Union (IoU) provide an essential measure of overall performance, a qualitative analysis offers deeper insights into how different training strategies affect the final segmentation output. Figure 7 provides a visual comparison of predicted masks for various building structures, contrasting the four primary training scenarios: training from scratch without and with data augmentation, and transfer learning without and with data augmentation. This visual evidence reinforces the quantitative findings presented in the previous sections.

Scratch Non-DA (Training from Scratch without Data Augmentation): This method consistently shows the poorest performance. Predicted masks are often fragmented, exhibit significant under-segmentation, and have poor boundary definition. This approach struggles with both simple and complex building layouts, frequently missing entire building structures, as seen in rows 2, 4, and 6 of Figure 7. The poor performance is a direct result of the model's inability to generalize from limited training data, leading to a high degree of overfitting.



Scratch DA (Training from Scratch with Data Augmentation): The introduction of data augmentation leads to a substantial improvement over the non-augmented scratch training. The predicted masks show better building detection and more complete segmentations. The boundary definition is improved, although it remains inferior to transfer learning approaches.

TL Non-DA (Transfer Learning without Data Augmentation): Leveraging pre-trained ImageNet weights results in a significant performance gain over all scratch-based methods. Predicted masks are much more accurate in both building detection and boundary delineation. The models generally capture building shapes well, but some noise and over-segmentation issues, such as connecting separate buildings, can still be observed.

TL DA (Transfer Learning with Data Augmentation): The combination of transfer learning and data augmentation consistently delivers the best overall performance. This method produces the most accurate segmentations that closely match the ground truth, with clean boundary detection and minimal noise. The models show consistent and reliable performance on various building types and densities, effectively handling complex urban layouts, as seen in rows 2 and 5 of Figure 7.



Fig. 7. Comparison of predicted masks from the four training methods scenarios, including training from scratch and transfer learning without and with data augmentation

4. Discussion

4.1. Impact of Training Strategies on Model Performance and Generalization

Our findings demonstrate the critical role of transfer learning (TL) and data augmentation (DA) in enhancing model performance and generalization for building extraction. The consistent and significant increase in IoU scores—generally 10-15% higher when employing transfer learning compared to training from scratch—highlights the immense value of leveraging pre-trained weights. This suggests that the generalized feature representations learned from large datasets like ImageNet provide a robust foundation, accelerating convergence and improving the models' ability to recognize fundamental visual patterns relevant to buildings.

While data augmentation provided modest individual IoU gains of 1-3% on the more homogeneous WHU dataset, its true significance was revealed on the challenging Vietnamese UAV Building Dataset. On this complex real-world data, DA's primary strength lies in its ability to mitigate severe overfitting, which is a direct consequence of high data variability and limited samples. As detailed in Section 3.2, overfitting

percentages for models trained from scratch on the Vietnamese dataset ranged from 21% to 30%, but with the introduction of DA, these rates consistently dropped to a range of 8% to 14%. By expanding data diversity through geometric transformations, models are prevented from merely memorizing training instances, leading to more robust and generalizable feature learning.

The effect of combining transfer learning and data augmentation resulted in the highest IoU scores and the lowest overfitting rates across all models and both datasets. On the Vietnamese building dataset, this combined approach reduced overfitting to a remarkable range of 3% to 7%, representing a substantial improvement over various scenarios. This combined approach acts as a powerful regularization mechanism, where pre-trained knowledge guides the network, and augmented data refines this knowledge to the specific nuances of the building domain. This is crucial for real-world applications where data diversity is high but sample size might be limited.

4.2. Architectural Strengths and Limitations for Building Extraction

The comparative analysis of model architectures revealed distinct strengths and weaknesses. DeepLabV3+ consistently emerged as a strong performer across various training scenarios on both datasets, achieving high IoU scores (e.g., 0.877 on WHU and a leading 0.858 on the Vietnamese dataset for the optimal configuration) while maintaining a relatively moderate parameter count (around 12.3 million). This suggests an efficient design capable of extracting rich contextual information vital for accurate segmentation.

Transformer-based models like SegFormer also demonstrated strong performance, particularly in transfer learning scenarios, reaching competitive IoU scores (e.g., 0.871 on WHU and 0.855 on the Vietnamese dataset for the optimal configuration) and indicating the growing potential of attention mechanisms for capturing global context in building extraction.

However, their computational overhead might be a consideration for resource-constrained environments. Conversely, Unet consistently underperformed relative to its parameter count, suggesting it may not be the optimal choice for these datasets, especially when trained from scratch. Its simpler skip connections might be less effective at handling the complex variations and occlusions present in the datasets compared to more advanced architectures like Unet++ or DeepLabV3+ that offer richer feature fusion or multi-scale contextual aggregation. This underscores that architectural design, rather than just raw parameter count, is a pivotal determinant of performance. Our findings indicate no direct linear correlation between model complexity (parameter count) and IoU performance, reinforcing the importance of architectural efficiency in achieving high accuracy.

4.3. Addressing Real-World Dataset Challenges

The significant performance gap observed between the WHU Building Dataset and the Vietnamese UAV Building Dataset confirms the difficulty of real-world UAV imagery. As shown in Fig. 5, all models consistently achieved lower IoU scores on the Vietnamese dataset compared to the WHU dataset. This persistent performance gap serves as the most direct evidence that the Vietnamese dataset is more challenging. The increased difficulty can be attributed to specific real-world complexities that are less prevalent in the more homogeneous WHU dataset, such as high architectural diversity, frequent occlusions, spectral variability, and inconsistencies from varied collection campaigns. This emphasizes that models performing well on benchmark datasets may still struggle with the complexities of diverse real-world data. Despite this inherent difficulty, the study demonstrates that robust methodological approaches, particularly the combined application of transfer learning and data augmentation, are highly effective in helping models generalize and perform reliably.

4.4. Practical Implications and Future Directions

The findings of this study offer several practical implications for building extraction in remote sensing. The clear benefits of combining transfer learning with data augmentation provide a strong recommendation for developing robust and generalizable models, especially when working with limited or

highly diverse real-world datasets. For practical deployment, architectures like DeepLabV3+ and SegFormer, which demonstrated high accuracy with moderate parameter counts (around 11-12 million parameters), represent compelling choices for balancing performance with computational constraints. Future work could explore the impact of more advanced data augmentation techniques, such as neural style transfer or adversarial training, to further enhance robustness against unseen variations. Investigating the performance of these models on even larger and more diverse real-world datasets from different geographical regions would also provide valuable insights. Additionally, exploring lightweight model architectures or knowledge distillation techniques could further optimize the balance between accuracy and computational efficiency for deployment on edge devices.

5. Conclusion

This study comprehensively evaluated the performance of six advanced deep learning architectures for building extraction, focusing on the critical roles of transfer learning and data augmentation across both a benchmark (WHU Building Dataset) and a challenging real-world (Vietnamese UAV Building Dataset) scenario.

Our research yielded several significant findings. Transfer learning and data augmentation are necessary strategies for achieving high performance and mitigating overfitting in semantic segmentation tasks, particularly for diverse and complex real-world imagery. The synergistic combination of these two techniques consistently yielded the highest Intersection over Union (IoU) scores and the lowest overfitting rates across all investigated models and datasets. For instance, on the Vietnamese dataset, this combined approach reduced overfitting from a baseline of up to 30% (scratch training) to a remarkable 3-7%. The study also highlights that architectural efficiency is a more crucial determinant of performance than raw parameter count. Models like DeepLabV3+ and SegFormer consistently achieved top-tier IoU scores (e.g., 0.858 and 0.855 on the Vietnamese dataset, respectively) with moderate parameter counts (11-12 million), demonstrating their efficiency. Furthermore, the Vietnamese UAV Building Dataset proved significantly more challenging than the WHU benchmark, as evidenced by consistently lower IoU scores across all models and training conditions (Fig. 5).

These findings provide recommendations for developing robust building extraction models. The combined application of transfer learning and data augmentation is the optimal training paradigm for achieving high accuracy and generalization in real-world scenarios, especially with limited data. For practical deployment, architectures like DeepLabV3+ and SegFormer are highly suitable choices, offering an excellent balance between high accuracy and computational efficiency. These models can be effectively utilized for automated building mapping in urban planning, disaster response, and environmental monitoring.

However, this study was conducted using a specific set of six deep learning architectures and two distinct building datasets. While these datasets are diverse, they may not capture the full spectrum of global building complexities. Additionally, the computational resources utilized were provided by Google Colaboratory's free-tier, which can have dynamic availability and resource limitations, potentially influencing the scale of experiments.

Future research can build upon these insights by exploring more advanced data augmentation techniques, such as neural style transfer or adversarial training, to further enhance robustness against unseen variations. Investigating the performance of these models on even larger and more diverse real-world datasets from different geographical regions would also provide valuable insights. Finally, developing and evaluating lightweight model architectures or knowledge distillation techniques could further optimize the balance between accuracy and computational efficiency for deployment on edge devices or in resource-constrained environments.

Acknowledgment

This research was sponsored by the Vietnam Ministry of Training and Education under the project B2024-MDA-09.

References

- Chen, H.; Lu, S.(2019). Building extraction from remote sensing images using SegNet. 2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC),
- Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV),
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition,
- Fayaz, S.; Ahmad Shah, S.Z.; ud din, N.M.; Gul, N.; Assad, A.J.R.a.i.c.s.; & communications. (2024). Advancements in data augmentation and transfer learning: A comprehensive survey to address data scarcity challenges. *17*(8), 14-35.
- Freire, S.; Santos, T.; Navarro, A.; Soares, F.; Silva, J.; Afonso, N.;...Sensing, R. (2014). Introducing mapping standards in the quality assessment of buildings extracted from very high resolution satellite imagery. *90*, 1-9.
- Ji, S.; Wei, S.; Lu, M.J.I.T.o.g.; sensing, r. (2018). Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *57*(1), 574-586.
- Lafarge, F.; Descombes, X.; Zerubia, J.; Pierrot-Deseilligny, M.J.I.J.o.p.; sensing, r. (2008). Automatic building extraction from DEMs using an object approach and application to the 3D-city modeling. *63*(3), 365-381.
- Li, H.; Xiong, P.; An, J.; Wang, L.J.a.p.a. (2018). Pyramid attention network for semantic segmentation.
- Li, Q.; Mou, L.; Sun, Y.; Hua, Y.; Shi, Y.; Zhu, X.X.J.I.T.o.G.; Sensing, R. (2024). A review of building extraction from remote sensing imagery: Geometrical structures and semantic attributes. *62*, 1-15.
- Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. (2017). Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition,
- Liu, Z.; Shi, Q.; Ou, J.J.I.T.o.G.; Sensing, R. (2022). LCS: A collaborative optimization framework of vector extraction and semantic segmentation for building extraction. *60*, 1-15.
- Luo, L.; Li, P.; Yan, X.J.E. (2021). Deep learning-based building extraction from remote sensing images: A comprehensive review. *14*(23), 7982.
- Luo, M.; Ji, S.; Wei, S.J.I.J.o.S.T.i.A.E.O.; Sensing, R. (2023). A diverse large-scale building dataset and a novel plug-and-play domain generalization method for building extraction. *16*, 4122-4138.
- Marastoni, N.; Giacobazzi, R.; Dalla Preda, M.J.J.o.C.V.; Techniques, H. (2021). Data augmentation and transfer learning to classify malware images in a deep learning context. *17*(4), 279-297.
- Mayer, H. (1999). Automatic object extraction from aerial imagery—a survey focusing on buildings. *74*(2), 138-149.
- Ronneberger, O.; Fischer, P.; Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention.
- Sahar, L.; Muthukumar, S.; French, S. P.J.I.T.o.G.; Sensing, R. (2010). Using aerial imagery and GIS in automated building footprint extraction and shape recognition for earthquake risk assessment of urban inventories. *48*(9), 3511-3520.
- Wang, L.; Fang, S.; Meng, X.; Li, R.J.I.T.o.G.; Sensing, R. (2022). Building extraction with vision transformer. *60*, 1-11.
- Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P.J.A.i.n.i.p.s. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *34*, 12077-12090.
- Yang, H.L.; Yuan, J.; Lunga, D.; Laverdiere, M.; Rose, A.; Bhaduri, B.J.I.J.o.S.T.i.A.E.O.; Sensing, R. (2018). Building extraction at scale using convolutional neural network: Mapping of the united states. *11*(8), 2600-2614.

- Zhang, T.; Tang, H.J.R.S. (2018). A comprehensive evaluation of approaches for built-up area extraction from landsat oli images using massive samples. *II*(1), 2.
- Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. International workshop on deep learning in medical image analysis,