Journal of Mining and Earth Sciences

Website: https://jmes.humg.edu.vn

Cross-domain transferability of deep learning models for building extraction from UAV imagery: a case study and benchmark in Vietnam

Dung Trung Pham *

Hanoi University of Mining and Geology, Hanoi, Vietnam

ARTICLE INFO

Article history:
Received 20th Jan. 2026
Revised 08th Apr. 2026
Accepted 23rd Apr. 2026

Keywords:
Building extraction,
Domain gap,
DeepLabV3+,
UAV imagery,
Urban morphology,
Vietnam.

ABSTRACT

Automated building extraction from Unmanned Aerial Vehicle (UAV) imagery plays an important role in urban analysis and large-scale geospatial applications. However, models trained on international datasets often struggle when deployed in regions with distinct local urban morphologies. This study evaluates the cross-domain transferability of such models in the Vietnamese urban context. Rather than testing domain adaptation methods, the paper focuses on assessing the performance degradation of source-trained models relative to a localized baseline.

Using the DeepLabV3+ architecture with a ResNet-18 backbone, models trained on representative global datasets were evaluated against the HUMG benchmark derived from UAV imagery in Vietnam. The results reveal a pronounced domain gap, with cross-domain Intersection over Union (IoU) values decreasing substantially and reaching as low as 0.642. Error analysis indicates that this degradation is primarily driven by the prevalence of high-density "tube houses," which frequently cause instance merging and omission errors in non-localized models.

In contrast, a model trained in local 10 cm resolution imagery achieved an IoU of 0.857. An efficiency plateau was also observed, whereby approximately 2,000 training samples (40% of the dataset) were sufficient to achieve near-optimal performance at very high resolution. To contextualize these findings within geospatial mapping practice, a simplified geometric model was introduced to explore the potential relationship between IoU and map-scale requirements. However, it is emphasized that overlap-based metrics provide only an indicative assessment and are insufficient to demonstrate regulatory compliance or full geometric validity. Overall, the study highlights the necessity of localized data for reliable building extraction in complex urban environments such as Vietnam.

*Corresponding author
E - mail: phamtrungdung@humg.edu.vn
DOI: 10.46326/JMES.2026.67(4).03

1. Introduction

Automated building extraction from UAV and Very-High-Resolution (VHR) satellite imagery has become a core component of modern urban management, cadastral mapping, and rooftop solar monitoring (Nguyen et al., 2025). Early studies in Vietnam urban landscapes utilized object-based approaches and morphological filtering to delineate commercial and residential structures in dense areas like Ho Chi Minh City (Vu et al., 2005). While these foundational methods highlighted the challenges of spectral similarity in crowded residential blocks, modern Deep Learning models have since achieved more remarkable success in semantic segmentation. However, their performance remains heavily dependent on the representativeness and quality of the training samples (Tuia et al., 2016). The process of constructing pixel-wise annotated datasets is not only labor-intensive but also significantly costly in terms of time and human resources (Iqbal & Ali, 2020; Yao et al., 2021). This challenge is further exacerbated when deploying pre-trained models in new geographical regions, where variations in sensors, illumination conditions, and architectural appearances lead to a "domain shift," severely degrading the model's generalization capabilities (Dias et al., 2022; Yao et al., 2021). When such models are applied without local calibration, their reliability in practical mapping scenarios can be significantly reduced.

Distribution inconsistency between the source and target domains currently represents the primary bottleneck in remote sensing (Peng et al., 2022). To address this, various Domain Adaptation (DA) techniques have been proposed, including subspace alignment (Song et al., 2019), decoupling style and semantic features (Chen et al., 2024), and curriculum learning to capture urban idiosyncrasies such as the size and spatial relations of buildings (Zhang et al., 2019; Zhang et al., 2017). Recent advances have also introduced anomaly detection-based frameworks to handle distribution bias (Zhao et al., 2023) or multitarget DA through domain-common approximation learning (Zhang et al., 2024) and dual-contrast adaptation coupling global context with geometry information (Xu et al., 2025). Although these methods have shown promising results under specific experimental conditions, their effectiveness depends strongly on assumptions regarding domain similarity and the availability of auxiliary data.

However, the urban landscape in Vietnam presents unique morphological challenges that existing DA frameworks have yet to resolve fully. Deng et al. (2023) highlighted that Southeast Asian architecture, characterized by high-density rural and urban settlements, requires high adaptability from deep learning models. In particular, the widespread presence of narrow "tube houses" in dense urban clusters creates complex spatial configurations with minimal inter-building spacing, which often results in geometric ambiguity. In the absence of localized training data, models frequently suffer from severe "instance merging" errors where multiple narrow, interconnected buildings are incorrectly segmented as a single large polygon (Hu et al., 2024).

Rather than proposing or evaluating domain adaptation methods, this study addresses a complementary and practically motivated problem by assessing the cross-domain transferability of deep learning-based building extraction models in the Vietnamese context. Specifically, the research investigates (i) the extent of performance degradation when models trained on international datasets are directly applied to Vietnam, and (ii) how this performance compares with a locally trained baseline derived from high-resolution UAV imagery. In addition, the study examines how spatial resolution interacts with training sample size to influence extraction accuracy under local conditions.

Through systematic experiments on the HUMG Benchmark, this research yields several critical findings. We demonstrate that models trained on international data experience a substantial performance drop when applied in Vietnam, for example, with the WHU dataset exhibiting an IoU reduction from 0.868 in self-domain evaluation to 0.642 in cross-domain deployment. In contrast, models trained using localized data at a 10 cm spatial resolution achieve an improved IoU of 0.857. An efficiency plateau is further identified at approximately 2,000 training samples, indicating that near-optimal performance can be obtained without exhaustive manual annotation. These findings provide empirical evidence of the limitations of direct model transfer and highlight the importance of localized training data for reliable

building extraction in complex urban environments such as those in Vietnam.

2. Method

2.1. Dataset description and experimental design

The experimental framework utilizes two primary data groups: an international collection for cross-domain evaluation and the local HUMG benchmark. This framework is designed to evaluate source-to-target transferability. Data partitioning follows strict protocols to ensure scientific rigor and prevent spatial leakage. All training and validation samples are geographically independent, ensuring that specific geographic scenes do not appear in more than one subset. As detailed in Table 1, validation set sizes remain constant across all training scales to provide an unbiased performance comparison. Training subsets are partitioned into incremental intervals of 20, 40, 60, 80, and 100 percent to identify the precise data saturation point for building extraction.

The HUMG dataset originally featured a 10 cm Ground Sampling Distance. To enable a fair comparison with international datasets under equivalent resolution conditions, a 30 cm version was generated using a specific downsampling protocol. The bilinear interpolation method is applied to downsample the original 10 cm UAV imagery to 30 cm resolution. This method is used to preserve spectral characteristics while reducing spatial detail to meet international standards. Corresponding building masks were resampled using the Nearest Neighbor method to preserve the integrity of binary labels and prevent blurred boundaries in the ground truth. This 30 cm dataset serves as the target domain in Group B experiments to assess the domain gap under identical resolution settings (Table 1) and (Table 2).

The evaluation process consists of three distinct phases. Group A conducts internal validation to determine the performance potential of each original dataset under consistent geographic and sensor conditions. Group B performs cross-domain tests by applying existing international models to the HUMG benchmark to quantify performance degradation caused by the domain gap. Group C undertakes localized optimization of the HUMG data to establish the optimal sample size and evaluate potential compliance with topography mapping standards in Vietnam.

The international datasets exhibit varying degrees of morphological similarity to the Vietnamese context (Figure 1). The WHU dataset contains large residential buildings with regular geometries and exhibits low similarity due to the contrast between widely spaced Western villas and the dense interconnected tube houses in Vietnam.

The Japan scenario represents a modern high-density environment but features more regulated planning and homogeneous roofing textures compared to the diverse multi-colored roofs in the HUMG dataset. In contrast, the Thailand dataset demonstrates high similarity to HUMG as both regions share Southeast Asian architectural features, including spontaneous urban growth and comparable building scales. The HUMG benchmark is notable for its unique tube houses, which are narrow, deep structures constructed wall to wall. These buildings form continuous blocks with complex spectral signatures from varied roofing materials, presenting a significant challenge for models trained strictly on international data.

The HUMG benchmark includes approximately 40,000 buildings sampled from diverse urban, rural, and industrial areas across Vietnam. Aerial imagery was captured using a DJI Phantom 4 RTK platform under clear sky conditions and high solar elevation between 10:00 AM and 2:00 PM to minimize shadows. The UAV was operated at a consistent altitude to maintain a Ground Sampling Distance of 10 cm, ensuring sufficient spatial detail to delineate narrow boundaries. A team of experts performed pixel-wise digitization in QGIS following a rigorous multi-stage quality control protocol. This systematic workflow establishes a high-fidelity ground truth standard specifically tailored to evaluate model generalization in the Vietnamese context.

2.2. Model architecture and transfer learning

The building extraction framework in this study is built upon the DeepLabV3+ architecture (Chen et al., 2017), which integrates an encoder-decoder structure designed to capture multi-scale contextual information while maintaining precise spatial details. To enhance the model's generalization and convergence speed, especially under constrained data regimes, Transfer Learning

S1 (WHU) 512 x 512 pixel.

S2 (Japan) 1024 x 1024 pixel.

S3 (Thailand) 1024 x 1024 pixel.

S4 (Mozambique) 1024 x 1024 pixel.

S5 (Kenya) 1024 x 1024 pixel.

S8 (HUMG) 512 x 512 pixel.

Figure 1. Representative building morphologies and spectral characteristics across the global datasets and the localized S8 (HUMG) benchmark.

Table 1. Configuration of global building datasets for cross-domain evaluation.

Scenario	Training	Validation	Image size	Resolution
S1 (WHU)	4736	2416	512x512	30 cm
S2 (Japan)	614	153	1024x1024	30 cm
S3 (Thailand)	664	166	1024x1024	30 cm
S4 (Mozambique)	418	105	1024x1024	30 cm
S5 (Kenya)	2394	596	1024x1024	30 cm
S6 (Ja_Thai)	1278	319	1024x1024	30 cm
S7 (Ja_Thai_Mo)	1696	424	1024x1024	30 cm

Table 2. Summary of localized HUMG dataset configurations across multiple training scales and resolutions. The image size is 512x512 pixels.

Scenario	Training scale	Training samples	Validation samples	Image size	Resolution (GSD)
S1A	20%	1056	1545	512 x 512	10 cm
S2A	40%	2119	1545	512 x 512	10 cm
S3A	60%	3180	1545	512 x 512	10 cm
S4A	80%	4243	1545	512 x 512	10 cm
S5A	100%	5445	1545	512 x 512	10 cm
S1B	20%	143	198	512 x 512	30 cm
S2B	40%	286	198	512 x 512	30 cm
S3B	60%	429	198	512 x 512	30 cm
S4B	80%	592	198	512 x 512	30 cm
S5B	100%	754	198	512 x 512	30 cm

is employed by utilizing a pre-trained ResNet-18 backbone as the encoder (He et al., 2016). This approach leverages hierarchical features previously learned from the ImageNet dataset, providing a significantly higher accuracy compared to training from scratch (Zhuang et al., 2020). The ResNet-18 backbone, characterized by 11.7 million parameters, is chosen for its computational efficiency and its deep residual learning framework, which effectively mitigates the vanishing gradient problem through "shortcut connections." Formally, for an input x and a desired underlying mapping H(x), the network learns the residual function F(x) = H(x) - x, producing the output Y = F(x) + x. This architecture ensures that if the weights of F(x) reduce to zero, the block performs an identity mapping, facilitating the robust training of deep networks (He et al., 2016).

Central to the encoder's performance is the Atrous Spatial Pyramid Pooling (ASPP) module (He et al., 2015), which utilizes parallel atrous (dilated) convolutions at different rates. This configuration allows the model to capture features at multiple scales without increasing parameter counts or losing spatial resolution, a feature critical for maintaining robustness when building sizes vary significantly within the urban landscape (Liu et al., 2019). The decoder branch complements the encoder by recovering fine-grained spatial details through the fusion of high-level semantic features with low-level spatial features from the backbone. By concatenating the upsampled encoder output with corresponding low-level features, the model refines building boundaries and ensures the generation of precise segmentation masks. This integrated approach, combining a residual encoder with an advanced spatial pyramid decoder, enables accurate Intersection over Union (IoU) assessment even within the complex and dense urban environments characterized by the HUMG and global datasets.

2.3. Model training and experimental setup

The building extraction model was trained using the PyTorch framework with the DeepLabV3+ architecture and a ResNet-18 backbone (Figure 2). To enhance model generalization, we applied a systematic data augmentation strategy using the Albumentations library, focusing on geometric transformations to simulate diverse flight orientations.

Specific parameters included ShiftScaleRotate with limits of 5% for shifting, 10% for scaling, and 15^0 for rotation. Horizontal and vertical flips were also implemented with application probabilities of 0.5 and 0.3, respectively. Optimization was performed using the Adam optimizer with an initial learning rate of 10-4. We utilized a StepLR scheduler with a step size of 10 and a gamma of 0.75, alongside

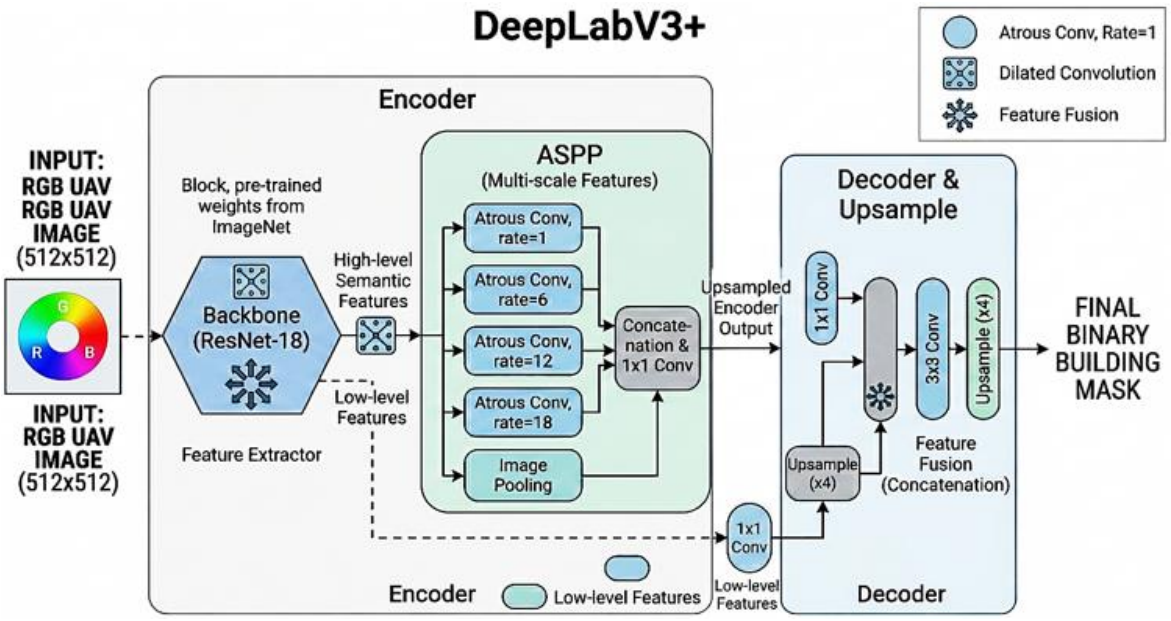

Figure 2. DeepLabV3+ with backbone ResNet-18 architecture.

the BCEWithLogitsLoss function. The training process employed a batch size of 8 and was set for a maximum of 50 epochs. To prevent overfitting while maintaining computational efficiency, early stopping was applied with a patience of 10 epochs and a minimum delta of 10^{-4} (Table 3).

Table 3. Summary of model hyperparameters, optimization settings, and data augmentation configurations.

	(Hyperparameters)	Configuration (Value)
Optimization	Optimizer/ Initial LR	Adam/10^{-4}
	LR Scheduler	StepLR (Step:10, γ:0.75)
	Loss Function	BCE With Logits Loss
Training	Batch Size/ Max Epochs	8/50
	Early Stopping	Patience: 10, Δmin: 0.0001
Geometric	Shift/Scale/ Rotate	±5%/ ±10%/±15^0
	Flips (H/V)	p=0.5/p=0.3

Given the computational complexity of evaluating multiple international scenarios and training scales, this study focuses on identifying primary performance trends through a controlled single run for each configuration. This approach serves as an exploratory baseline to quantify the domain gap and establish the overall behavior of data saturation across different spatial resolutions. The final model selection was based on the best validation checkpoint to ensure that the reported metrics represent the peak generalization capability of each specific training regime.

2.4. Mathematical relationship between extraction accuracy (IoU) and topographic mapping standards

To evaluate the potential of deep learning-based building extraction for professional surveying, this study establishes a theoretical relationship between the Intersection over Union metric and the planimetric error standards defined by national regulations in Circular 07/2021/TT-BTNMT. We utilize a simplified geometric model assuming a rectangular building of width (W) and length (L) extracted with a consistent boundary displacement error (Δs). In this assessment, we follow geodetic principles where the maximum allowable error is defined as twice the standard planimetric error (2Δs). Under this assumption, the Intersection and Union areas are formulated as follows:

$$Area(Intersection) = (W - \Delta s) \times (L - \Delta s) \quad (1)$$

$$Area(Union) = (L \times W) + \big(\Delta s \times (L + W)\big) - (\Delta s^2) \quad (2)$$

Consequently, the minimum required IoU threshold to satisfy a specific planimetric error Δs is derived as:

$$IoU = \frac{(W - \Delta s) \times (L - \Delta s)}{(L \times W) + \big(\Delta s \times (L + W)\big) - (\Delta s^2)} \quad (3)$$

This geometric derivation provides a baseline for assessing how pixel-wise overlap metrics correspond to traditional positional accuracy requirements. Based on the derived equation, Table 4 summarizes the IoU thresholds required to meet legal accuracy standards for various mapping scales.

Table 4. Correlation between mapping scales, positional accuracy, and required iou thresholds.

Map Scale	Max allowable Error (Δs)	Standard Error (Δs)	Min IoU (10x10m House)	Min IoU (5x10m Tube House)
1:1,000	0.5 m	0.25 m	90%	86%
1:2,000	1.0 m	0.50 m	82%	75%
1:5,000	2.5 m	1.25 m	60%	47%

We define the standard planimetric error as half of the maximum allowable error for each scale. We compare a standard 10 by 10 m building and a 5 by 10 m tube house to reflect the unique urban morphology of Vietnam. While this model offers useful practical insights, we acknowledge that actual cartographic compliance requires additional independent evaluation of boundary shape and topological consistency.

3. Results

To provide a comprehensive evaluation of building extraction performance, five widely recognized quantitative metrics were employed: Intersection over Union (IoU), F1-Score, Precision, Recall, and Overall Accuracy. These metrics are

standard in the field of semantic segmentation and allow for a multifaceted analysis of model reliability, encompassing both pixel-level classification and geometric boundary fidelity. While IoU serves as the primary indicator for spatial overlap, the inclusion of Precision and Recall is essential for identifying specific error types, such as false positives (over-segmentation) and false negatives (omissions) in dense urban blocks. The use of these five metrics for benchmarking remote sensing models is well-established in the literature (Fernandez-Moral et al., 2018; Padilla et al., 2020).

3.1. Performance analysis of global datasets (Self-test Evaluation)

The internal reliability of global datasets was first established through a self-test protocol (as detailed in Table 5). The experiment results show that the WHU dataset attained the highest performance metrics among the tested groups, achieving an IoU of 0.868 and an F1-Score of 0.904. This high performance, characterized by balanced Precision of 0.921 and Recall of 0.912, provides the quantitative baseline for building feature convergence (Figure 3). In contrast, datasets from other regions showed lower internal consistency. The Kenya (S5) and Mozambique (S4) datasets yielded significantly lower IoU scores of 0.521 and 0.650, respectively. A systematic error analysis reveals that these discrepancies are closely linked to roof complexity and environmental contrast. Notably, the Recall deficit in the Kenya dataset (0.610) compared to its Precision (0.795) confirms that models trained on sparse or low-contrast architectural patterns struggle to detect nearly 40% of building footprints even within their own domain.

This suggests that high spectral similarity between roofing materials and the surrounding soil leads to significant omission errors.

Multi-source scenarios, such as Ja_Thai_Mo (S7), maintained a stable IoU of 0.728, demonstrating the impact of regional aggregation on model capacity. However, even with regional aggregation, these models still struggle with the unique morphological complexity of dense residential subsets found in the Vietnamese context.

3.2. Evaluation of cross-domain adaptability and the domain gap

To evaluate the generalization capability of global datasets within the Vietnamese urban context, models trained on international data were tested directly against the HUMG test sets at both 10 cm and 30 cm resolutions (as illustrated in Figure 4 and 5). The quantitative results, presented in Table 5 and 6, reveal a significant performance degradation when compared to the localized S8 (HUMG) dataset, which serves as the "upper bound" with an IoU of 0.857 (10 cm) and 0.814 (30 cm).

The "domain gap" is most evident in models trained on geographically distant regions such as Kenya (S5) and Mozambique (S4), where IoU scores plummeted below 0.52. Even the high-standard WHU (S1) dataset, despite its excellent internal performance (self-test IoU of 0.868), only achieved an IoU of 0.642 when applied in Vietnam. This substantial decline (0.868÷0.642) highlights that high internal accuracy does not guarantee cross-domain robustness, particularly when faced with distinct regional architectural styles. A notable exception among single-source models is Thailand (S3), which reached an IoU of 0.749 (at 10 cm).

This suggests that architectural similarities in Southeast Asian urban morphology and roofing materials facilitate better feature transfer. While multi-source scenarios like Ja_Thai (S6) and Ja_Thai_Mo (S7) improved stability (IoU > 0.72),

Table 5. Cross-domain performance metrics on HUMG 10 cm test set (including S1-S8).

Scenario	IoU	F1-Score	Precision	Recall	Accuracy
S1 (WHU)	0.642	0.769	0.856	0.723	0.932
S2 (Japan)	0.690	0.810	0.870	0.772	0.870
S3 (Thailand)	0.749	0.851	0.867	0.875	0.836
S4 (Mozambique)	0.458	0.605	0.886	0.489	0.876
S5 (Kenya)	0.521	0.676	0.795	0.610	0.795
S6 (Ja_Thai)	0.731	0.836	0.882	0.811	0.943
S7 (Ja_Thai_Mo)	0.722	0.829	0.867	0.812	0.946
S8 (HUMG)	0.857	0.921	0.931	0.915	0.974

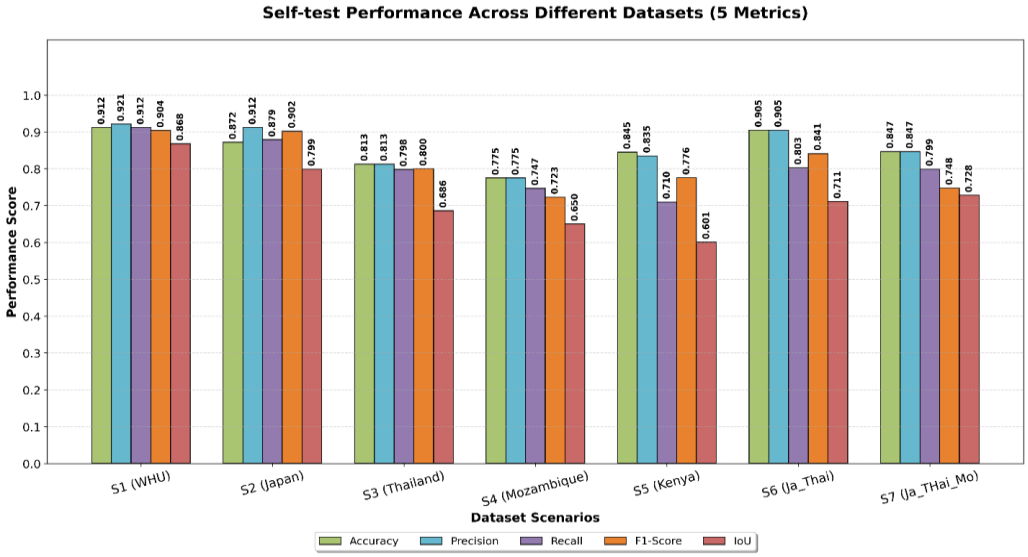

Figure 3. Self-test performance metrics (Accuracy, Precision, Recall, F1-Score, and IoU) across global scenarios S1-S7 at 30 cm GSD.

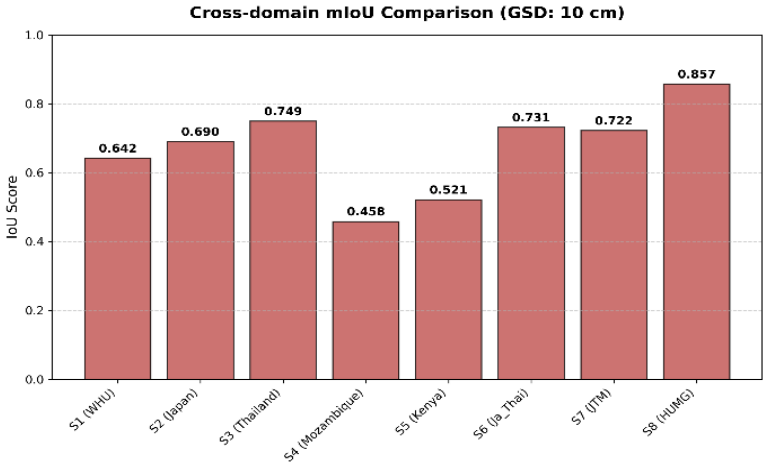

Figure 4. Comparative bar chart of IoU scores for all scenarios, highlighting the gap between global datasets and the S8 baseline in the case of 10 cm GSD.

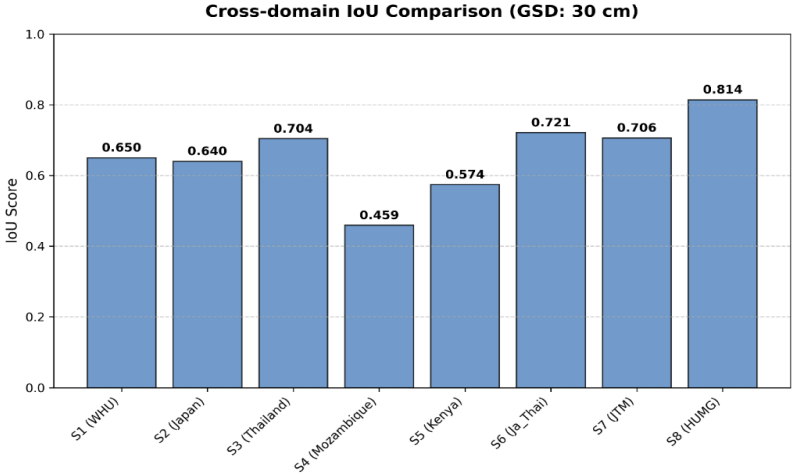

Figure 5. Comparative bar chart of IoU scores for all scenarios, highlighting the gap between global datasets and the S8 baseline in the case of 30 cm GSD.

they still remained 13÷15% lower in IoU compared to the localized S8 model.

A fine-grained error analysis reveals that the disparity in both Precision and Recall between the global models and the S8 baseline is driven by specific morphological and environmental factors. Specifically, the high-density "tube house" structures in Vietnam, characterized by shared walls and narrow gaps, create complex spatial relations that global models often fail to resolve. In the HUMG 10 cm test set (Table 5), the Recall for S4 (0.489) and S5 (0.610) was significantly lower than their Precision (0.886 and 0.795, respectively), indicating a high rate of false negatives. This suggests that low spectral contrast and complex roof textures lead to significant omission errors (false negatives), whereas the "instance merging" effect occurs when adjacent buildings are merged into a single polygon. These findings provide empirical data on the performance of international datasets when applied to cadastral and urban contexts in Vietnam.

The qualitative results presented in Figure 6 and Figure 7 visually confirm the "domain gap" identified in the quantitative analysis.

Table 6. Cross-domain performance metrics on HUMG 30 cm test set (including S1-S8).

Scenario	IoU	F1-Score	Precision	Recall	Accuracy
S1 (WHU)	0.650	0.780	0.861	0.724	0.959
S2 (Japan)	0.640	0.770	0.829	0.747	0.829
S3 (Thailand)	0.704	0.822	0.799	0.853	0.799
S4 (Mozambique)	0.459	0.622	0.913	0.482	0.936
S5 (Kenya)	0.574	0.726	0.835	0.654	0.846
S6 (Ja_Thai)	0.721	0.834	0.840	0.832	0.965
S7 (Ja_Thai_Mo)	0.706	0.825	0.883	0.779	0.965
S8 (HUMG)	0.814	0.897	0.901	0.893	0.978

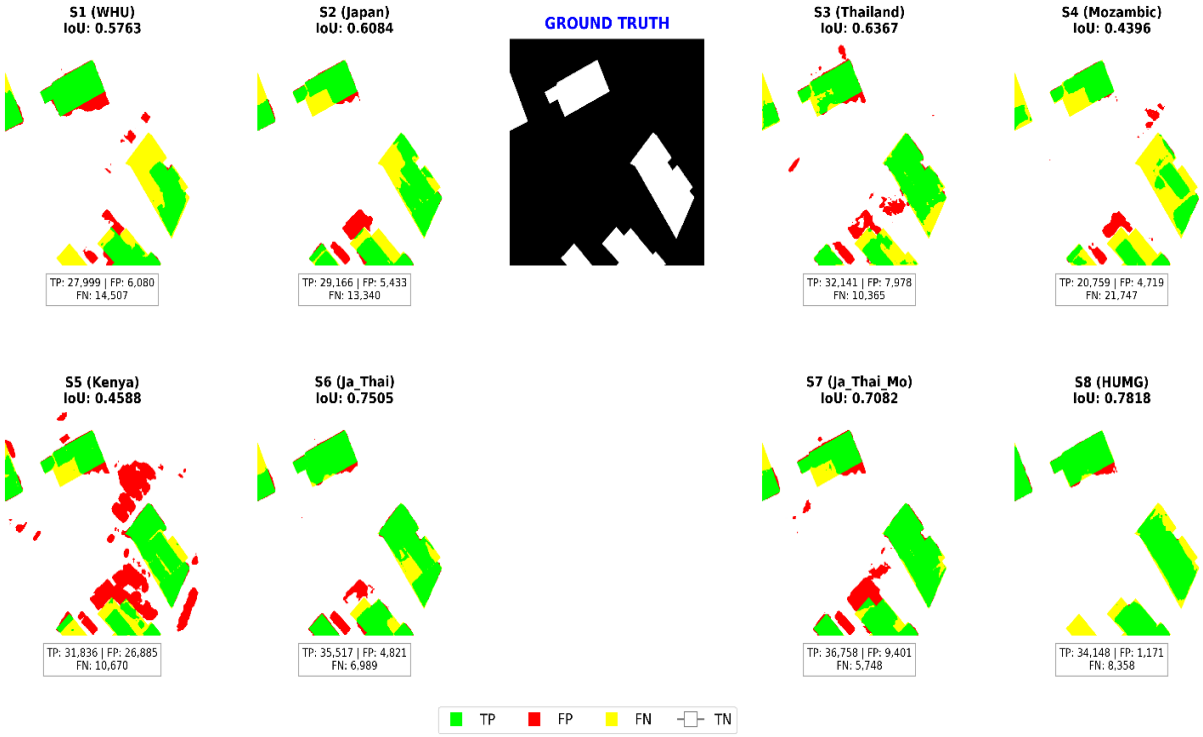

Figure 6. Qualitative comparison of building extraction results on the HUMG 10 cm test set using models trained on global datasets versus the localized HUMG baseline.

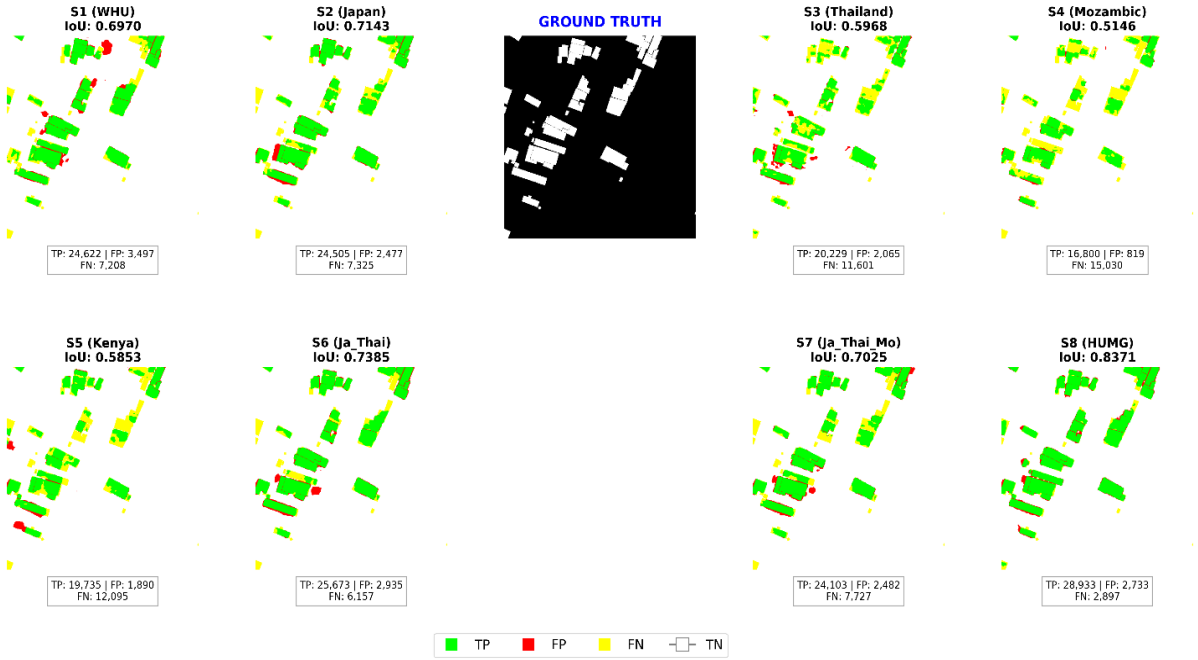

Figure 7. Qualitative comparison of building extraction results on the HUMG 30-cm test set using models trained on global datasets versus the localized HUMG baseline.

Models trained on global datasets, particularly S1 (WHU) and S5 (Kenya), struggle to delineate the high-density "tube house" structures characteristic of the HUMG dataset.

As illustrated in Figure 6, the S1 model frequently suffers from "instance merging" errors, where multiple narrow, interconnected buildings are incorrectly segmented as a single large polygon due to the lack of similar morphological features in its training domain. Furthermore, Figure 7 highlights significant false negative rates in models from developing contexts like S5, where low spectral contrast between roofing materials and the surrounding environment leads to the omission of nearly 40% of building footprints. In contrast, the localized S8 (HUMG) model demonstrates superior edge definition and boundary precision, successfully identifying individual structures even within complex, wall-to-wall urban blocks.

3.3. The necessity of the HUMG benchmark and optimization of training scale

The development of the localized HUMG dataset (S8) recorded a peak IoU of 0.857 at 10 cm GSD and 0.814 at 30 cm GSD. The precision (0.931) and recall (0.915) values obtained at the 10 cm resolution indicate the model's capacity to identify complex and narrow building structures common in the Vietnamese urban context, while minimizing instances of false positives and false negatives.

Through a systematic trend analysis across varying training scales, a critical finding of this study lies in the relationship between spatial resolution and the training data scale. As illustrated in the performance trends, the two resolutions exhibit distinct behaviors regarding data saturation:

For the 10 cm GSD imagery, the model reaches a performance plateau at an early stage of data increment (Figure 8). At 40% of the training scale (S2A), the IoU reached 0.854, compared to 0.857 at

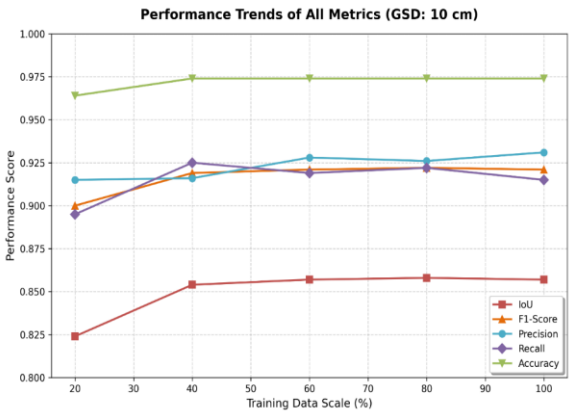

Figure 8. Comparative analysis of performance saturation: IoU trends with 10 cm GSD relative to training data scale.

the 100% scale (S5A). The performance metrics across varying scales for 10 cm GSD are detailed in Table 7, showing that the increase in IoU from 40÷100% training data is 0.003.

Conversely, the 30 cm dataset shows a linear growth trend, reaching its maximum IoU of 0.814 at the 100% training scale (S5B).

Table 7. Performance metrics of building extraction using 10 cm GSD imagery across varying training data scales (20÷100 %).

Scenario	IoU	F1-Score	Precision	Recall	Accuracy
S1A (20%)	0.824	0.900	0.915	0.895	0.964
S2A (40%)	0.854	0.919	0.916	0.925	0.974
S3A (60%)	0.857	0.921	0.928	0.919	0.974
S4A (80%)	0.858	0.922	0.926	0.922	0.974
S5A (100%)	0.857	0.921	0.931	0.915	0.974

Even at maximum capacity, the IoU for 30 cm GSD remains 4.3% lower than the 10 cm GSD counterpart at the same scale (Table 8).

The relationship between spatial resolution and data saturation is visually documented in Figure 9 and Figure 10.

Table 8. Performance metrics of building extraction using 30 cm GSD imagery across varying training data scales (20÷100%).

Scenario	IoU	F1-Score	Precision	Recall	Accuracy
S1B (20%)	0.743	0.851	0.857	0.848	0.967
S2B (40%)	0.774	0.871	0.876	0.869	0.972
S3B (60%)	0.798	0.887	0.892	0.882	0.975
S4B (80%)	0.802	0.89	0.899	0.881	0.976
S5B (100%)	0.814	0.897	0.901	0.893	0.978

For 10 cm GSD imagery, Figure 9 shows that the extraction quality at a 40% training scale (S2A) is visually similar to the results at 100% (S5A). In contrast, as shown in Figure 10, the 30 cm GSD models require the full 100% training scale (S5B) to achieve stabilized boundary definitions.

However, the 30 cm outputs exhibit different geometric characteristics compared to the 10 cm results, particularly regarding edge sharpness (Figure 11).

These findings provide a quantitative basis for selecting 10 cm GSD as a benchmark for high-precision urban monitoring.

Figure 9. Qualitative comparison of building extraction results on the HUMG 10 cm test set using a localized HUMG dataset with varying training data scales.

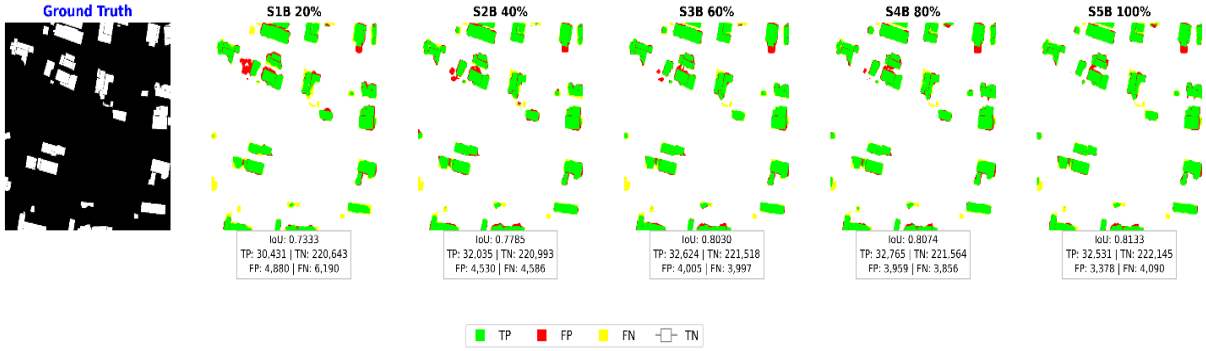

Figure 10. Qualitative comparison of building extraction results on the HUMG 30 cm test set using a localized HUMG dataset with varying training data scales.

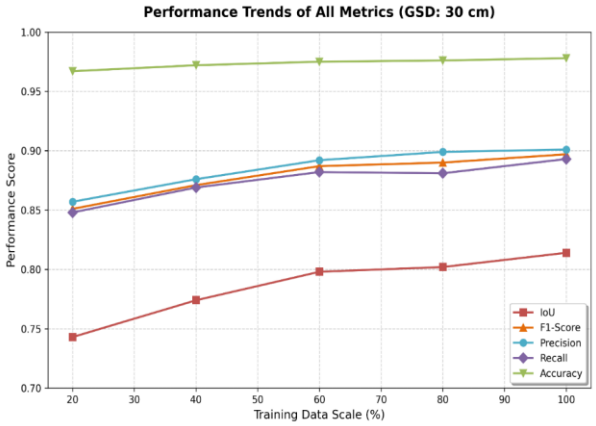

Figure 11. Comparative analysis of performance saturation: IoU trends with 30 cm GSD relative to training data scale.

4. Discussion

4.1. Evaluation of global datasets and the "Domain Gap" challenge

The experimental results reveal a significant performance degradation, commonly referred to as the "domain gap," when applying models trained on international datasets to the Vietnamese urban context (Tuia et al., 2016). Even the WHU dataset, which is widely regarded as a gold standard with a self-test IoU of 0.868, saw its performance plummet to 0.642 when deployed on the HUMG test set. This sharp decline (0.868÷0.642) is highly consistent with the findings of (Dias et al., 2022) and (Zhao et al., 2023), who argued that distribution bias between source and target domains is the primary barrier for deep learning models in real-world environments.

A systematic multi-factor error analysis reveals that this discrepancy stems primarily from differences in urban morphology. Global datasets often represent Western-style architecture with clear spacing. In contrast, Vietnamese cities are characterized by high-density "tube houses" built wall-to-wall. Consistent with (Zhang et al., 2017), our results confirm that "urban idiosyncrasies" regarding scale and spatial relations are decisive factors. International models frequently suffer from "instance merging" errors, incorrectly segmenting narrow, interconnected buildings as a single large polygon. Furthermore, datasets from low-contrast environments failed to distinguish rooftops from road surfaces, supporting the concerns of (Yao et al., 2021) regarding the high entropy and uncertainty in cross-domain building segmentation.

4.2. Assessment of the HUMG dataset and UAV development strategies

The development of the localized HUMG dataset proved to be a decisive factor, achieving a superior IoU of 0.857. This result significantly complements and extends the works of (Deng et al., 2023) and (Nguyen et al., 2019). While Deng et al. (2023) focused on rural buildings at broader scales, the present study highlights the effectiveness of 10 cm resolution imagery for complex urban residential blocks, in an overlap-based accuracy sense, where narrow building boundaries and adjacency relationships are particularly challenging (Hu et al., 2024).

An important finding of this study is the identification of an efficiency plateau at approximately 2,000 training samples for 10 cm GSD imagery. At this scale, the model rapidly converges on key morphological characteristics of local building structures, and further increases in training data yield only marginal improvements in IoU (approximately 0.3%), despite a substantial increase in annotation effort. This observation provides a practical guideline for optimizing manual labeling resources by prioritizing data resolution and representativeness over sheer dataset size (Iqbal & Ali, 2020).

The results also reveal a discrepancy between empirical performance and the expectations of recent multi-source domain adaptation (MSDA) theories (Peng et al., 2019; Xu et al., 2025; Yao et al.,

2021). While MSDA frameworks aim to extract invariant features from multiple global datasets, the findings here indicate that such approaches may still be insufficient when local urban morphology differs substantially from the source domains. The structural complexity of Vietnamese "tube houses" appears to limit the transferability of supposedly invariant features, suggesting that localized training data remain essential for reliable performance in such environments.

4.3. Analysis of compliance with geodetic and cadastral mapping standards

Based on Circular No. 07/2021/TT-BTNMT and the theoretical correlations summarized in Table 4, a conceptual relationship can be established between allowable positional error, standard planimetric error, and the minimum IoU required under different mapping scales. The simplified geometric model suggests that, for smaller-scale maps, lower IoU thresholds may still be theoretically consistent with generalized positional accuracy requirements, whereas larger-scale mapping demands substantially higher overlap accuracy, particularly for narrow building typologies such as tube houses.

However, it is important to emphasize that this analysis remains indicative and theoretical in nature. The derived IoU thresholds are based on simplified geometric assumptions and reflect pixel-level overlap rather than true positional accuracy, boundary shape fidelity, or topological correctness. As such, they should not be interpreted as direct evidence of compliance with official geodetic or cadastral mapping standards.

In practice, deep-learning-generated building masks often exhibit softened or irregular boundaries, instance-merging effects, and topological inconsistencies, particularly for small and closely spaced structures Vu et al. (2005). Therefore, even when overlap-based metrics such as IoU exceed certain indicative thresholds, rigorous post-processing steps remain essential. These include independent positional accuracy assessment, boundary regularization, topology validation, and cartographic refinement before any operational use in official land administration or large-scale mapping workflows. The results of this study thus reinforce the view that VHR imagery and localized training data are necessary but not sufficient conditions for professional-grade geospatial data production (Nguyen et al., 2025).

5. Conclusions

This study presents a systematic evaluation of the cross-domain transferability of deep learning-based building extraction models by comparing widely used international datasets with the localized HUMG benchmark in the Vietnamese context. The experimental results clearly confirm the presence of a substantial domain gap between global datasets and local urban morphology. Significant degradation in Intersection over Union (IoU) is observed when models trained on international data are deployed in Vietnam, demonstrating that distinctive characteristics such as high-density tube-house architecture and heterogeneous roofing materials induce distribution shifts that current global models are not yet able to handle reliably. Detailed error analysis further reveals that instance merging remains the dominant failure mode, as models struggle to separate narrowly spaced and interconnected residential structures.

The development of the localized HUMG dataset using 10 cm ground sampling distance imagery significantly improves model performance, achieving an IoU of 0.857 under local testing conditions. This result indicates a clear advantage of localized training data in preserving overlap-based boundary accuracy and mitigating common cross-domain errors such as omissions and object merging. However, while these results represent a substantial improvement over direct cross-domain deployment, they should be interpreted as an indicative performance assessment rather than conclusive evidence of geometric or regulatory compliance. In particular, for large-scale mapping applications (e.g., 1:1,000 and 1:2,000), automated extraction outputs still require comprehensive post-processing procedures before they can be considered suitable for professional cartographic use.

An additional contribution of this study is the identification of an efficiency plateau at ultra-high spatial resolution. The experiments demonstrate that approximately 40% of the training data, corresponding to around 2,000 samples, is sufficient to reach near-optimal IoU performance, whereas substantially increasing the annotation effort yields

only marginal accuracy gains. This finding highlights the importance of data quality and spatial resolution over dataset size and provides a practical guideline for optimizing manual labeling resources in UAV-based building extraction workflows.

Despite the encouraging results, several limitations remain. The current approach is affected by rounding artifacts and softened building corners, which reduce geometric fidelity in complex urban settings. Future research will focus on expanding the dataset to additional geographic regions and investigating semi-supervised learning strategies to further reduce annotation costs. Moreover, integrating polygon regularization techniques with boundary-sensitive metrics such as the BF1 score will be essential to enable a more rigorous evaluation of geometric integrity and to better align automated extraction results with the stringent requirements of official surveying and land administration systems.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Gemini to improve the English language and grammar. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the published article.

Author contributions

The author was solely responsible for all aspects of this study, including the conceptualization of the research framework, dataset construction, experimental design, implementation of deep learning models, and quantitative and qualitative analysis of the results. The author also conducted the interpretation of findings, prepared all figures and tables, and wrote, revised, and approved the final manuscript.

References

Chen, J., Zhu, J., He, P., Guo, Y., Hong, L., Yang, Y., & Sensing, R. (2024). Unsupervised domain adaptation for building extraction of high-resolution remote sensing imagery based on decoupling style and semantic features. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-17.

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv. https://doi.org/10.48550/arXiv.1706.05587.

Deng, X., Liang, Y., Li, X., & Xu, W. (2023). Recognition and spatial distribution of rural buildings in Vietnam. *Land*, 12(12), 2142. https://doi.org/10.3390/land12122142.

Dias, P., Tian, Y., Newsam, S., Tsaris, A., Hinkle, J., & Lunga, D. (2022). Model assumptions and data characteristics: Impacts on domain adaptation in building segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-18. https://doi.org/10.1109/TGRS.2022.3160000.

Fernandez-Moral, E., Martins, R., Wolf, D., & Rives, P. (2018). A new metric for evaluating semantic segmentation: Leveraging global and contour accuracy. 2018 *IEEE Intelligent Vehicles Symposium (IV),* 1041-1046. https://doi.org/10.1109/IVS.2018.8500539

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9), 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

Hu, D., Gan, V. J., & Zhai, R. (2024). Automated BIM-to-scan point cloud semantic segmentation using a domain adaptation network with hybrid attention and whitening (DawNet). *Advanced Engineering Informatics,* 164, 105473. https://doi.org/10.1016/j.aei.2024.105473.

Iqbal, J., & Ali, M. (2020). Weakly-supervised domain adaptation for built-up region segmentation in aerial and satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167, 263-275. https://doi.org/10.1016/j.isprsjprs.2020.07.001.

Liu, Y., Gross, L., Li, Z., Li, X., Fan, X., & Qi, W. (2019). Automatic building extraction on high-resolution remote sensing imagery using deep

convolutional encoder-decoder with spatial pyramid pooling. *IEEE Access*, 7, 128774-128786. https://doi.org/10.1109/ACCESS.2019.2940527.

Nguyen, A., Luu, H., Phan, A., Bui, H., & Nguyen, T. (2019). Cau Giay: A dataset for very dense building extraction from Google Earth imagery. *2019 6th NAFOSTED Conference on Information and Computer Science (NICS),* 401-406. https://doi.org/10.1109/NICS48861.2019.9023812.

Nguyen, H. N., Dung, T. T. T., Hung, T. D., & Tuan, H. M. (2025). Deep learning-based segmentation of rooftop photovoltaic panels in Ho Chi Minh City, Vietnam. *2025 17th International Conference on Knowledge and System Engineering (KSE).*

Padilla, R., Netto, S. L., & Da Silva, E. A. (2020). A survey on performance metrics for object-detection algorithms. *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 237-242. https://doi.org/10.1109/IWSSIP48289.2020.9145130.

Peng, J., Huang, Y., Sun, W., Chen, N., Ning, Y., & Du, Q. (2022). Domain adaptation in remote sensing image classification: A survey. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 9842-9859. https://doi.org/10.1109/JSTARS.2022.3221375.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., & Wang, B. (2019). Moment matching for multi-source domain adaptation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),* 713-722.

Song, S., Yu, H., Miao, Z., Zhang, Q., Lin, Y., & Wang, S. (2019). Domain adaptation for convolutional neural networks-based remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 16(8), 1324-1328. https://doi.org/10.1109/LGRS.2019.2895714.

Tuia, D., Persello, C., & Bruzzone, L. (2016). Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2), 41-57. https://doi.org/10.1109/MGRS.2016.2548504.

Vu, T. T., Matsuoka, M., & Yamazaki, F. (2005). Object-based extraction of buildings from very high resolution satellite images: A case study of Ho Chi Minh City, Viet Nam. *Proceedings of the 26th Asian Conference on Remote Sensing (ACRS)*, 7-11.

Xu, G., Deng, M., Zhu, J., Sun, G., Gou, Z., Guo, Y., ... & Sensing, R. (2025). A dual-contrast adaptation network coupling global context and geometry information for cross domain building extraction. *IEEE Transactions on Geoscience and Remote Sensing*.

Yao, X., Wang, Y., Wu, Y., & Liang, Z. (2021). Weakly-supervised domain adaptation with adversarial entropy for building segmentation in cross-domain aerial imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 8407-8418. https://doi.org/10.1109/JSTARS.2021. 3104505.

Zhang, F., Liu, K., Liu, Y., Wang, C., Zhou, W., Zhang, H., ... & Sensing, R. (2024). Multitarget domain adaptation building instance extraction of remote sensing imagery with domain-common approximation learning. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1-16.

Zhang, Y., David, P., & Gong, B. (2017). Curriculum domain adaptation for semantic segmentation of urban scenes. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2020-2030.

Zhang, Y., David, P., Foroosh, H., & Gong, B. (2019). A curriculum domain adaptation approach to the semantic segmentation of urban scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8), 1823-1841. https://doi.org/10.1109/TPAMI.2019.2902392.

Zhao, S., Zhou, X., & Hou, D. (2023). An anomaly detection-based domain adaptation framework for cross-domain building extraction from remote sensing images. *Applied Sciences*, 13(3), 1674. https://doi.org/10.3390/app13031674.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., ... & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43-76.