Remote Sensing Applications: Society and Environment 41 (2026) 101876

Contents lists available at ScienceDirect

Remote Sensing Applications: Society and Environment

journal homepage: www.elsevier.com/locate/rsase

Optimising deep learning for building extraction: Dataset efficiency and model backbones under data constraints

Dung T. Pham [a], Thuong V. Tran [b,*], Xuan Zhu [b], Hung N. Pham [c]

[a] Faculty of Geomatics and Land Administration, Hanoi University of Mining and Geology, Hanoi, 10000, Viet Nam
[b] School of Earth, Atmosphere and Environment, Monash University, Clayton, VIC, 3800, Australia
[c] Phenikaa School of Computing, Phenikaa University, Hanoi, 10000, Viet Nam

ARTICLE INFO

Keywords:
Building extraction
Dataset optimisation
Transfer learning
Data augmentation
Overfitting mitigation
DeepLabV3+
ResNet

ABSTRACT

Accurate building extraction from very high-resolution (VHR) satellite imagery is critical for urban planning, disaster response, and environmental monitoring. However, the performance of deep learning models remains highly sensitive to training sample size, model complexity, and learning strategy, especially in data-scarce scenarios. Here, we systematically evaluated the DeepLabV3+ architecture with ResNet backbones (i.e., ResNet-18, 50, 101, 152) across varying training sample sizes (i.e., 20–120 % of the primary WHU dataset) and three training approaches: random initialisation, ImageNet pre-training, and pre-training with data augmentation. Our results revealed a U-shaped relationship between dataset size and performance, with segmentation accuracy peaking at full dataset usage and declining when additional samples introduce redundancy. Peak performance reached 88.0 % IoU and 93.2 % F1-score under optimised configuration, while shallower models achieved optimal performance under limited data availability. We found that hybrid learning strategies are essential for mitigating overfitting and achieving high accuracy, with transfer learning improving accuracy by 7.9 % IoU (5.8 % F1-score), and data augmentation offering an additional 1–3 % IoU (0.2–2.3 % F1-score) gain in low-data settings ($\leq$40 %). Deeper networks (ResNet-101/152) achieved superior performance only when trained with $\geq$60 % of the dataset and appropriate regularisation. The stability of these data-dependent model selection patterns was further confirmed through external validation on the Japan Building Dataset, demonstrating transferability across geographic contexts. Our findings yield practical and generalisable guidelines: (i) avoiding unnecessary dataset expansion, (ii) prioritising transfer learning and augmentation when data is scarce, and (iii) aligning model depth with data availability. By explicitly linking model selection to data availability, this study supports efficient and reliable deployment of deep learning for urban analytics and disaster response under realistic annotation constraints.

1. Introduction

Accurate extraction of building footprints from remotely sensed data underpins a wide range of geospatial applications, including

* Corresponding author. School of Earth, Atmosphere and Environment, Monash University, 9 Rainforest Walk Room 251, Clayton Campus, Australia.

E-mail addresses: phamtrungdung@humg.edu.vn (D.T. Pham), thuong.tran@monash.edu (T.V. Tran), xuan.zhu@monash.edu (X. Zhu), hung.phamngoc@phenikaa-uni.edu.vn (H.N. Pham).

https://doi.org/10.1016/j.rsase.2026.101876
Received 10 July 2025; Received in revised form 24 December 2025; Accepted 7 January 2026
Available online 9 January 2026

urban development monitoring (Ayala et al., 2021), population estimation, disaster risk reduction (Ghaffarian et al., 2018), infrastructure planning (Patel and Shukla, 2025), and cadastral updating (Vincent M and P, 2024). In rapidly urbanising regions, particularly where informal settlements proliferate without formal records, reliable building maps are essential for equitable service delivery and effective urban governance. Moreover, building footprints serve as foundational inputs for higher-level tasks such as land-use classification, change detection, and emergency response (Chen et al., 2023). However, automated extraction from very high-resolution (VHR) imagery remains challenging due to rooftop variability, occlusion from vegetation and shadows, and acquisition inconsistencies (Fernandez-Moral et al., 2018; Liu et al., 2018). Traditional approaches based on hand-crafted features, spectral thresholds, and rule-based classifiers have proven inadequate in term of generalisability (Li et al., 2019). While object-based image analysis (OBIA) improved segmentation capabilities, its dependence on extensive parameter tuning and sensitivity to illumination and scene variability limits its scalability (Hay and Castilla, 2006; Ye et al., 2023). Deep learning, particularly Convolutional Neural Networks (CNNs), has since transformed the landscape of semantic segmentation by enabling end-to-end feature learning that generalises across diverse urban morphologies (Koldasbayeva et al., 2024).

Among data sources for urban object detection, VHR satellite imagery offers an ideal platform, capable of capturing sub-meter details at frequent time intervals and customisable angles (Iannelli et al., 2014; Liu et al., 2018). Unlike medium-resolution imagery, where buildings often fall below the pixel scale, VHR data allows the capture of subtle textural and structural features critical for accurate delineation (Kumar and Bhardwaj, 2020). However, the manual annotation of VHR imagery is labour-intensive, particularly in dense urban cores, creating a persistent scarcity of high-quality training data (He et al., 2024). To address these challenges, state-of-the-art architectures (e.g., DeepLabV3+), which combine *atrous* spatial pyramid pooling (ASPP) and encoder–decoder structures, enable effective multi-scale context integration and boundary preservation, making them particularly well-suited for complex urban scenes (Cha et al., 2024). Notwithstanding these architectural developments, DeepLabV3+ performance remains strongly conditioned on the quantity and representativeness of annotated training data, particularly in VHR applications where sample acquisition is costly and spatial heterogeneity is high (Wang et al., 2021, 2024). In limited-data scenarios, common in VHR applications, deep models risk overfitting, degraded generalisation, and unstable training dynamics (Salman and Liu, 2019). Alternatively, transfer learning, where models are initialised with weights pre-trained on large-scale datasets (e.g., ImageNet), enables better generalisation in downstream tasks (Alnagashi et al., 2024; Cherti and Jitsev, 2022; Wei et al., 2024). Data augmentation, which synthetically expands training diversity, further improves robustness in low-sample regimes (Balestriero et al., 2022; Shorten and Khoshgoftaar, 2019). Additionally, regularisation techniques such as dropout, batch normalization, and L2 penalties, contribute to stabilise training (Cai et al., 2019; Santos and Papa, 2022; Srivastava et al., 2014), while lightweight network architectures have been proposed to minimise over-parameterisation and computational burden (Bai et al., 2024). Nevertheless, several fundamental questions remain unresolved: (i) what is the minimum dataset size required to achieve robust performance? (ii) How does model depth affect accuracy under data scarcity? And (iii) what combination of learning strategies yields the most stable results when annotated data is limited?

Empirical studies over the past decade have begun to explore the relationships between dataset size, model complexity, and training strategy in the context of building extraction. One consistent observation is that although segmentation accuracy generally improves with increasing data volume, the rate of improvement diminishes, and overfitting becomes a concern when training samples are limited (Al-Ani et al., 2025). Model architecture further moderates these effects: shallower networks (e.g., ResNet-18) often outperform deeper backbones (e.g., ResNet-152) under data-scarce conditions, highlighting the importance of aligning model complexity with data availability (Cha et al., 2024). Earlier work also suggested that data quality and relevance may outweigh sheer volume. For instance, Yan et al. (2017) identified a practical range of 15,000–30,000 samples to achieve robust segmentation, while Zhu et al. (2020) reported that high accuracy could be achieved using approximately 58 % of the WHU dataset rather than the full sample set. Furthermore, Abriha and Szabó (2023) demonstrated that carefully curated, domain-specific samples can yield high accuracy without large generic datasets. These findings indicate that optimal performance arises from balancing data availability, model complexity, and learning strategy rather than maximizing any single component in isolation.

Hybrid learning strategies have further enhanced segmentation performance. For example, combining transfer learning and augmentation improved IoU by 5 % on SAR data (Wangiyana et al., 2022) and by up to 13 % overall (Shan et al., 2025). Pre-trained U-Net/ResNet hybrids achieved 1.2–1.6 % gains (Ait El Asri et al., 2023), while domain-adaptive fine-tuning has produced accuracies as high as 96.4 % (Neupane et al., 2025; Prakash et al., 2022). Architecture innovations also contributed to performance improvements: MobileNetV3-based DeepLabV3+ variants achieve IoU values of approximately 0.80 at reduced computational cost (Zhao et al., 2025), and attention-based modules (i.e., DAMM and IEU-Net) improve boundary delineation and robustness in noisy conditions (Cha et al., 2024). Although lightweight models offer faster inference, deeper networks (e.g., ResNet-101, ResNet-152) generally attain higher accuracy (~0.92 IoU) but exhibit increased susceptibility to overfitting under limited-data conditions (Cha et al., 2024; Heryadi et al., 2020). Such evidence demonstrates that, when combined with appropriate transfer learning, data augmentation, and regularisation, DeepLabV3+ remains a flexible and effective solution for building extraction across varying data regimes. However, systematic analyses are still lacking on how training sample size, backbone depth, and learning strategies interact across diverse operational settings.

Our study aims to conduct a comprehensive evaluation of DeepLabV3+ using four ResNet backbones (i.e., ResNet-18, -50, −101, and −152). The evaluation is implemented through a series of controlled experiments in which training sample size and learning strategy are systematically varied while all other training settings are held constant. Model performance is assessed under varying training sample sizes (20–120 % of the WHU dataset) and three training strategies are compared: random initialisation, ImageNet pre-training, and ImageNet pre-training with data augmentation. Instead of proposing a new segmentation architecture, the study examines how data availability, network depth, and training strategy influence segmentation accuracy and generalisation for building

extraction. Our contributions are threefold: (i) quantifying the impact of dataset size on segmentation accuracy and identify thresholds of diminishing returns; (ii) examining how backbone depth influences segmentation accuracy and overfitting behaviour across different data regimes; (iii) comparing learning strategies to determine the most effective approach in low-, medium-, and high-data settings. In this study, the term "data-driven model selection" refers to the explicit use of training data availability as the primary criterion for selecting backbone depth and learning strategy, rather than relying on peak accuracy alone. Model selection is therefore guided by empirical evidence on segmentation accuracy, overfitting behaviour, and training stability across multiple data regimes, enabling practical recommendations under realistic annotation constraints. Performance differences are interpreted using standard segmentation metrics, early stopping behaviour, and independent cross-dataset validation, providing a direct methodological basis for addressing the stated research objectives. These analyses provide practical guidance for resource-efficient building extraction and clarify how deep learning models can be applied reliably when annotated data are limited.

2. Materials and methods

Our study evaluates the performance of DeepLabV3+ for building footprint extraction from very high-resolution (VHR) imagery under varying data availability conditions. The evaluation is based on a series of controlled experiments in which training sample size, backbone depth, and learning strategy are explicitly varied while all other training settings are kept identical. Specifically, we investigate: (i) the effect of training sample size on validation accuracy and overfitting, (ii) the influence of backbone depth on segmentation accuracy and generalisation across different data regimes, and (iii) the comparative effectiveness of three training strategies (i.e., random initialisation, transfer learning, and transfer learning with data augmentation). For the purpose of model selection, configurations are compared within each data regime using a combination of segmentation accuracy (IoU and F1-score), overfitting indicators derived from training–validation divergence, and convergence stability during training. The preferred configuration within each data regime is defined as the model that achieves high accuracy without exhibiting early overfitting or unstable optimisation behaviour. All experiments were conducted using the WHU building dataset (Ji et al., 2018), employing four ResNet backbones (i.e., ResNet-18, ResNet-50, ResNet-101, and ResNet-152), across six training regimes.

2.1. Dataset and experimental design

We utilised the WHU building dataset for model training, validation, and evaluation. The dataset was acquired from an urban region covering approximately 450 square kilometres in Christchurch, New Zealand and comprises 8189 image patches containing over 187,000 manually annotated building footprints. Each image has a spatial resolution of approximately 0.3 m and a standardised size of 512 × 512 pixels. The WHU dataset is publicly accessible through the New Zealand government's geospatial portal at https://data.linz.govt.nz. Following the established protocol described by (Ji et al., 2018), we adopted the original data split, consisting of 4736 images for training, 1036 for validation, and 2416 for testing.

To confirm the stability and transferability of the optimal learning strategies identified using the WHU Dataset, an independent geographical validation was performed using the Japan Building Dataset. This dataset comprises Very High-Resolution (VHR) satellite imagery with significantly different structural characteristics and building densities compared to WHU. The inclusion of this dataset ensures a rigorous test of the generalisability of the optimal rules regarding sample size and training strategies. Further details on this dataset are provided in Section S2 of the Supplementary Material. All training code and pre-trained models used in this study are publicly available at: https://github.com/trungdungtdct/Building-extraction/tree/main.

To assess the influence of training sample size on model performance and generalisation, we designed six experimental scenarios with incrementally increasing proportions of the training set (Table 1). Specifically, we used 20 %, 40 %, 60 %, 80 %, 100 %, and 120 % of the original training samples. In all cases, the validation dataset remained fixed at 1036 images to ensure consistency in performance evaluation. The sixth scenario (120 %) extends the training dataset by incorporating an additional 947 images drawn from the original test set, increasing the total number of training samples to 5683. This experimental design enables systematic examination of how training data volume influences validation accuracy and overfitting behaviour, providing insights into optimal data requirements for reliable building segmentation.

Table 1
Description of training and validation dataset configurations across six experimental scenarios with varying training sample sizes.

Training dataset		Validation dataset
% of the original sample	Number of samples	Number of samples
20 %	947	1036
40 %	1894	1036
60 %	2842	1036
80 %	3789	1036
100 %	4736	1036
120 %	5683	1036

2.2. DeepLabV3+ architecture with ResNet backbones

DeepLabV3+ is a state-of-the-art semantic segmentation architecture that combines an encoder–decoder framework with atrous convolutions to enable high-precision segmentation (W. Liu et al., 2019). It has proven particularly effective for building extraction tasks in remote sensing and aerial imagery, owing to its ability to integrate multi-scale contextual information while retaining fine spatial detail (Du et al., 2021). The encoder employs a ResNet backbone (e.g., ResNet-50 or ResNet-101) to extract hierarchical features representations at multiple spatial resolutions. A central component of the architecture is the Atrous Spatial Pyramid Pooling (ASPP) module (He et al., 2015), which applies parallel atrous convolutions with multiple dilation rates (e.g., 6, 12, 18 in this study). These dilation rates were fixed across all experiments to ensure architectural consistency. By enlarging the receptive field without reducing spatial resolution, allowing the network to detect buildings of varying sizes across complex urban environments. The use of residual connections in ResNet addresses the vanishing gradient problem by reformulating each layer as a residual function, defined as $H(x) = F(x) + x$, where $F(x)$ represents the learned residual mapping (He et al., 2016). This structure ensures stable gradient propagation and efficient feature learning, particularly in deeper networks.

The decoder recovers spatial details by fusing high-level semantic features from ASPP with low-level features from intermediate ResNet stages. Low-level features are taken from the corresponding intermediate ResNet block and reduced to 48 channels using 1×1 convolutions prior to fusion, following the standard DeepLabV3+ design. ASPP outputs are bilinearly upsampled by a factor of 4 to match the spatial resolution of the low-level features. The fused feature maps are refined using 3×3 convolutions, followed by a final bilinear up-sampling ($\times$ 4) to restore the original input resolution and produce pixel-accurate building masks with well-defined boundaries.

In this study, the standard X-ception backbone was replaced with ResNet variants (ResNet-18, -50, −101, and −152) to assess the trade-offs between model complexity, accuracy, and computational efficiency. All ResNet backbones were implemented using identical encoder–decoder connections and ASPP configurations, with backbone depth being the only architectural variable. Deeper models generally achieve higher representational capacity at the cost of increased computation and memory usage, while shallower networks offer faster inference and reduced resource requirements (Table 2).

To maintain reproducibility, no architectural components beyond backbone depth were modified across experiments. The ASPP module retained the same convolutional structure and dilation rates for all configurations, and decoder operations were fixed across backbones. Pre-trained ResNet weights (ImageNet) were used when transfer learning was enabled. These design choices ensure that observed performance differences can be attributed directly to backbone depth, training data availability, and learning strategy rather than architectural inconsistencies. The ResNet-based DeepLabV3+ architecture offers several advantages for building extraction (Du et al., 2021; W. Liu et al., 2019; Wang et al., 2024). Dilated convolutions enable robust detection across varying building scales, while skip connections preserve structural details, critical for boundary delineation in complex urban environments. Pre-trained ResNet weights facilitate generalization in data-limited settings, and residual connections improve optimisation stability without increasing parameter count. The encoder–decoder structure implemented in this study exactly corresponds to the architecture used in all experiments (Fig. 1). These characteristics make the ResNet-based DeepLabV3+ architecture well suited for automated building segmentation in geospatial applications.

2.3. Training procedures and model configurations

We applied the DeepLabV3+ architecture with a ResNet backbone across six experimental scenarios, each corresponding to a different training sample size as outlined previously. Three training strategies were employed: (i) training from scratch, (ii) training with ImageNet pre-trained weights, and (iii) training with ImageNet pre-trained weights combined with data augmentation. To investigate the impact of model complexity, we implemented four ResNet variants of increasing depth: ResNet-18, ResNet-50, ResNet-101, and ResNet-152. All experiments were conducted using the Google Colaboratory cloud-based platform, which provided access to GPU-accelerated computing resources. Specifically, we utilised an NVIDIA Tesla T4 GPU (16 GB of GDDR6 memory, 256-bit memory interface), offering sufficient computational power for training deep neural networks under varying data and architecture constraints. CUDA-enabled parallel processing was employed to accelerate model training and inference. The models were implemented using PyTorch version 1.11.0, which provides flexibility for defining custom architectures and optimising training workflows. A batch size of 8 was adopted to balance memory usage and training stability. For external validation experiment using the Japan Building Dataset (Section 3.5), the batch size was reduced to 2 to accommodate increased memory demands, while all other training parameters were kept unchanged. For optimisation, Stochastic Gradient Descent (SGD) was used with a cross-entropy loss function, which is appropriate for pixel-wise classification in semantic segmentation tasks. An initial learning rate of 0.0001 was selected to ensure gradual

Table 2
Computational cost (FLOPs) and number of trainable parameters for DeepLabV3+ models using ResNet-18, ResNet-50, ResNet-101, and ResNet-152 backbones.

Model	Backbones	FLOPs	Params (M)	Computational Efficiency
DeepLabV3+	ResNet-18	36.5641	12.3293	Fastest & lightest, ideal for edge devices
	ResNet-50	73.3382	26.6776	Balanced speed and accuracy, widely used
	ResNet-101	112.207	45.6697	Higher accuracy, but slower inference
	ResNet-152	151.095	61.3134	Heaviest, best for high-precision tasks

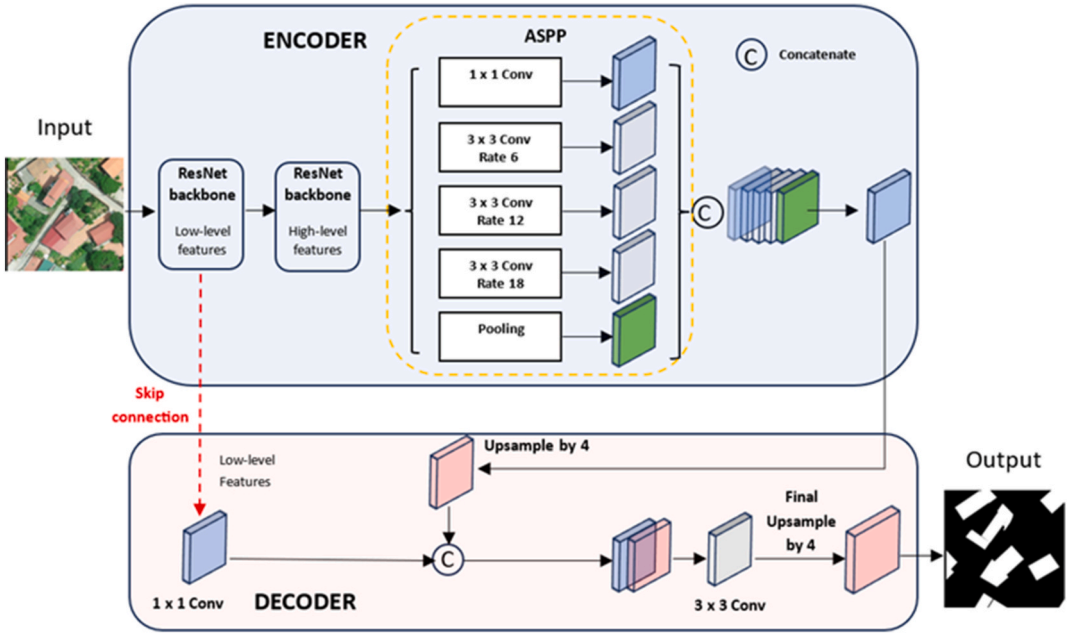

Fig. 1. Architecture of DeepLabV3+ used in this study for building footprint extraction. The encoder employs a ResNet backbone (ResNet-18, -50, −101, or −152) to extract multi-level features, followed by atrous spatial pyramid pooling (ASPP) to capture multi-scale contextual information. Low-level features are integrated into the decoder through skip connections to refine object boundaries. The architecture is identical across all experiments, with performance differences arising solely from variations in backbone depth, training data volume, and learning strategy.

convergence and stable weight updates. Each experiment was trained for 50 epochs, which was sufficient to ensure convergence across all backbone depths and training data scenarios. These settings were fixed across experiments to ensure comparability of results across architectures, data regimes, and learning strategies, while maintaining computational efficiency and reproducibility.

2.4. Training strategies

Transfer learning (TL) is a machine learning paradigm that enables knowledge acquired from a source domain to be transferred to improve performance on a related target task (Panigrahi et al., 2020). In building extraction, the source domain typically consists of large-scale datasets such as ImageNet (Deng et al., 2009), COCO (Lin et al., 2014), or Pascal VOC (Everingham et al., 2010), whereas the target domain focuses on semantic segmentation of buildings in remote sensing imagery (Gao et al., 2020). In this study, a TL was implemented by initialising the encoder with pre-trained ImageNet weights. Under this configuration, low-level visual features learned from natural images are retained, while high-level representations are fine-tuned for the building extraction task. Leveraging pre-trained ResNet backbones improves generalisation, faster convergence, and reduced overfitting, particularly under limited training data conditions (Bakirman et al., 2022; Ji et al., 2018). Compared to the models trained from scratch, the TL-based models consistently achieved higher accuracy while reducing training time and computational costs.

Data augmentation (DA) is also applied to artificially increase the size and diversity of the training dataset, thereby improving generalisation and robustness to real-world variability (Alomar et al., 2023). Exposure to a wider range of imaging conditions provides two primary benefits (Wangiyana et al., 2022): enhanced feature learning across buildings with varying scales, orientations, and illumination conditions, and reduced sensitivity to dataset-specific characteristics. Consequently, DA improves segmentation accuracy and training stability, particularly in remote sensing applications where annotated data are limited (Shorten and Khoshgoftaar, 2019).

A structured data augmentation pipeline was implemented to support systematic evaluation of building segmentation performance. The pipeline consists of three categories of transformations: input standardization (resizing to 512 × 512 pixels), geometric augmentation (ShiftScaleRotate, HorizontalFlip), and pixel-level enhancement (RandomBrightnessContrast, Blur, Sharpen, RGBShift). By combining these three types (i.e., standardization, geometric, and radiometric transformations), the pipeline balances data diversity with visual realism, enable models to generalise more effectively across heterogeneous urban environments while preserving structural integrity of building features.

2.5. Evaluation metrics

Intersection-over-Union (IoU), also referred to as the Jaccard Index, is a widely used evaluation metric in image segmentation, particularly for tasks involving semantic and instance segmentation (Fernandez-Moral et al., 2018). It quantifies the spatial overlap between predicted and ground-truth segmentation masks by computing the ratio of their intersection area to their union area.

Mathematically, IoU is defined as:

$$\text{IoU}(\%) = \frac{\text{Area of Overlap (Prediction} \cap \text{Ground Truth)}}{\text{Area of Union (Prediction} \cup \text{Ground Truth)}} \quad (1)$$

IoU values range from 0 (no overlap) to 1 (perfect alignment), providing an intuitive and interpretable measure of segmentation quality (Rezatofighi et al., 2019). Due to its strong correspondence with visual accuracy and its ability to penalize both false positives and false negatives, IoU has become the standard for evaluating pixel-wise classification performance in object detection and segmentation tasks (Guo et al., 2020; Yang et al., 2018; Zhang et al., 2020). In this study, IoU is used as the primary metric to evaluate the accuracy of building footprint extraction from aerial imagery.

IoU is particularly suitable for building segmentation for several reasons. First, accurate building extraction requires precise delineation of structural boundaries. Unlike pixel-wise accuracy metrics, which may inflate performance due to class imbalance, IoU directly evaluates the spatial alignment and penalizes both over-segmentation and under-segmentation errors (Yang et al., 2018). Second, aerial imagery typically contains buildings of varying sizes. Normalization via the union area enables fair comparisons across objects of different spatial scales, reducing bias toward larger structures. Third, urban scenes often exhibit strong background dominant, which can skew conventional accuracy metrics. By focusing exclusively on the overlap between predicted and reference building masks, IoU mitigates the effects of class imbalance and background pixels (Guo et al., 2020; Zhang et al., 2020). Finally, IoU aligns closely with human visual interpretation of segmentation quality. A high IoU value indicates strong geometric agreement between predicted building shapes and their true outlines, which is essential for applications such as urban planning, cadastral mapping, and post-disaster damage assessment, where geometric precision is critical. For these reasons, IoU provides a robust, geometrically meaningful, and practically relevant measure for evaluating building segmentation performance.

The F1-score, also commonly referred to as the Sørensen–Dice coefficient in segmentation and medical imaging contexts, offers a statistical measure of classification accuracy that balances two fundamental metrics: precision and recall (Luo et al., 2021). Precision quantifies the proportion of correct identifications $\left(TP/_{TP+FP}\right)$. Recall quantifies the proportion of actual positives that were correctly identified $\left(TP/_{TP+FN}\right)$. The terms TP, FP, and FN are the true positive, false positive, and false negative counts, respectively. The F1-score is calculated as the harmonic mean of these two metrics:

$$\text{F1} = 2\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

The F1-score ranges from 0 (worst) to 1 (best) and is particularly valuable because it penalizes models that achieve high precision at the expense of low recall, or vice-versa. In the context of building footprint extraction, models must often navigate a trade-off between maximizing the detection of all buildings (high recall) and ensuring that predicted masks are pure and do not include non-building areas (high precision). The F1-score provides a single metric that captures this trade-off in a statistically coherent manner. Similar to IoU, the F1-score inherently addresses the challenge of class imbalance by focusing on the positive class (building footprints) rather than overall pixel accuracy, which is often dominated by the background pixels. Accordingly, reporting both IoU and F1-score enables complementary evaluation of geometric accuracy and statistical classification performance, providing a comprehensive assessment of model behaviour in building extraction tasks.

3. Results

3.1. Effect of training sample size on segmentation accuracy

The validation accuracy (IoU) of DeepLabV3+ models was evaluated with four ResNet backbones across six training sample sizes (20 %–120 %) and three training strategies: random initialisation (RI), ImageNet pre-training (PT), and pre-training with data augmentation (PT + DA) (Table 3 and Fig. 2). Validation accuracy increased consistently as training sample size grews from 20 % to 100 %, regardless of backbone depth or training strategy. For example, with ResNet-18 using random initialisation, IoU improved from 72.0 % (20 %) to 83.2 % (100 %), while ResNet-50 improves from 73.0 % to 84.5 % (Table 3; Fig. 2a and b). Comparable trends were observed for the F1-score, detailed in Supplementary Table S1 and Fig. S1. When the training sample size exceeded 100 %, achieved through sample expansion to 120 %, performance began to decline across most configurations. A decline of approximately 2–4 % IoU was observed beyond this threshold, indicating diminishing returns and increased susceptibility to overfitting or redundant data (e.g., ResNet-18 dropped to 79.3 %; ResNet-152 decreased to 77.7 %).

With respect to learning strategy, models initialised with ImageNet pre-trained weights consistently outperformed those trained from scratch. For instance, at 100 % training data, ResNet-101 improved from 84.3 % (RI) to 87.4 % (PT), and further to 87.9 % with data augmentation (PT + DA) (Table 3 and Fig. 2c). The advantage of pre-training was most pronounced under limited-data conditions. At 20 % training size, ResNet-152 achieved an IoU of 68.8 % with random initialisation, compared to 74.9 % using PT and 78.0 % using PT + DA. These results demonstrate that prior knowledge and augmented sample diversity substantially enhance training stability and generalisation when annotated data is scarce (Fig. 2d). Gains associated with PT + DA diminished progressively as training data increased, suggesting reduced dependence on auxiliary strategies in data-rich regimes.

Backbone depth also influenced segmentation performance, with effects contingent on the training sample size and learning

Table 3
Impact of training sample size and training strategy on validation accuracy (IoU) for DeepLabV3+ with different ResNet backbones.

% of the original sample	Validation accuracy by IoU (%)											
	Random initialisation				Pre-trained ImageNet					Pre-trained ImageNet + DA		
	ResNet-18	ResNet-50	ResNet-101	ResNet-152	ResNet-18	ResNet-50	ResNet-101	ResNet-152	ResNet-18	ResNet-50	ResNet-101	ResNet-152
20 %	72.0	73.0	71.1	68.8	72.5	74.5	74.8	74.9	75.7	77.3	77.4	78.0
40 %	74.5	75.0	74.7	73.8	79.0	81.0	80.7	81.7	79.8	81.1	82.1	82.9
60 %	77.0	77.7	77.7	76.9	80.9	82.3	82.5	83.2	82.2	82.5	83.2	84.4
80 %	78.3	79.1	79.6	78.0	81.6	82.8	83.0	83.5	82.3	83.1	83.3	83.5
100 %	83.2	84.5	84.3	84.5	86.0	87.5	87.4	87.5	86.6	87.8	87.9	88.0
120 %	79.3	80.9	80.6	77.7	82.0	83.4	83.3	83.6	83.1	83.8	83.9	83.9

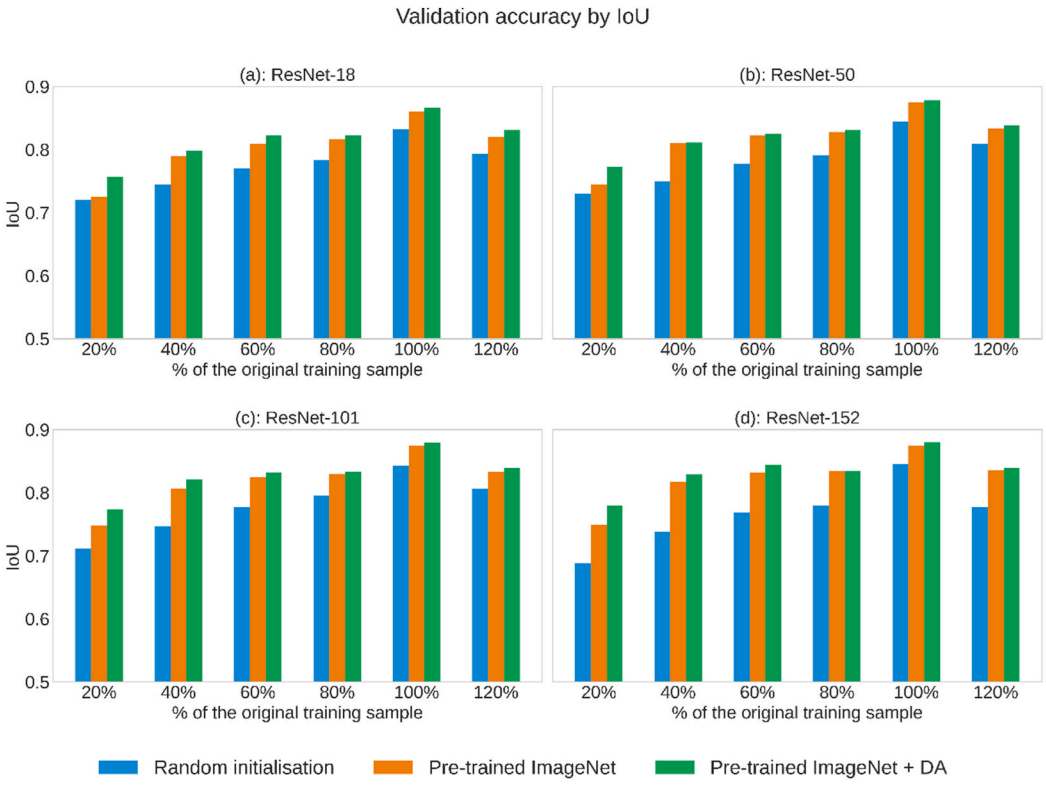

Fig. 2. Validation accuracy of DeepLabV3+ models with varying ResNet backbones under different training sample sizes. Each chart compares training approaches, including random initialisation, transfer learning, and transfer learning with data augmentation.

strategy. Under limited data conditions (20–40 %), shallower backbones such as ResNet-18 and ResNet-50 consistently outperformed deeper networks, especially when trained from scratch. For example, at 20 % training size, ResNet-50 achieved 73.0 % IoU, compared with 71.1 % for ResNet-101 and 68.8 % for ResNet-152. As training data increased, deeper architectures exhibited clear performance advantages. At 100 % training size with PT + DA, IoU values rose from 86.6 % (ResNet-18) to 88.0 % (ResNet-152), representing the highest segmentation accuracy achieved in this study. These results indicate that deeper backbones require sufficient training data and regularisation to realise their full representational capacity. Visual comparisons in Fig. 3 further corroborate these findings, showing progressive improvement in segmentation quality up to 100 % training size for ResNet-152 with PT + DA, followed by degradation at 120 %, consistent with increased overfitting or data redundancy.

3.2. Impact of learning strategies on segmentation accuracy

To assess the effectiveness of different training strategies, we evaluated improvements in segmentation accuracy (IoU) achieved through transfer learning and data augmentation across four ResNet backbones and varying training sample sizes (Table 4; Fig. 4). The strongest improvements were observed in intermediate regimes. For example, at 40 % of the original training data, ResNet-152 achieved the largest gain of 7.9 % IoU relative to random initialisation (Table 4; Fig. 4d), followed by ResNet-50 and ResNet-101, each with 6.0 % gains (Fig. 4b and c). Even the shallower ResNet-18 model benefited, with a 4.5 % improvement (Fig. 4a). These results highlight the effectiveness of ImageNet pre-trained, particularly under limited to moderate data availability. Comparable trends were observed for the F1-score, with substantial gains recorded across low-data regimes, most notably for ResNet-152 at 40 % training size (Table S2; Fig. S2), confirming the importance of pre-training for stable weight initialisation when annotated data are scarce.

Data augmentation further enhanced segmentation performance, particularly in low-data settings. At 20 % training size, augmentation contributed additional gains ranging from 2.6 % to 3.2 % IoU across the four architectures, with the largest improvement observed for ResNet-18 (Table 4; Fig. 4a–d). However, the effectiveness of augmentation diminished with increasing training size. For instance, beyond 40 % of the training data, augmentation-related gains dropped below 1.5 % IoU across all backbones and became negligible (<0.5 %) at 100 % and 120 % training sizes (Fig. 4). These patterns indicated that data augmentation is most effective when training samples are limited. Backbone depth also influenced the magnitude of improvement obtained from different learning strategies. Deeper architectures, such as ResNet-152 and ResNet-101, consistently showed larger gains from both pre-training and augmentation, particularly at 40 % and 60 % training sizes. In contrast, shallow models (e.g., ResNet-18) exhibited smaller but still

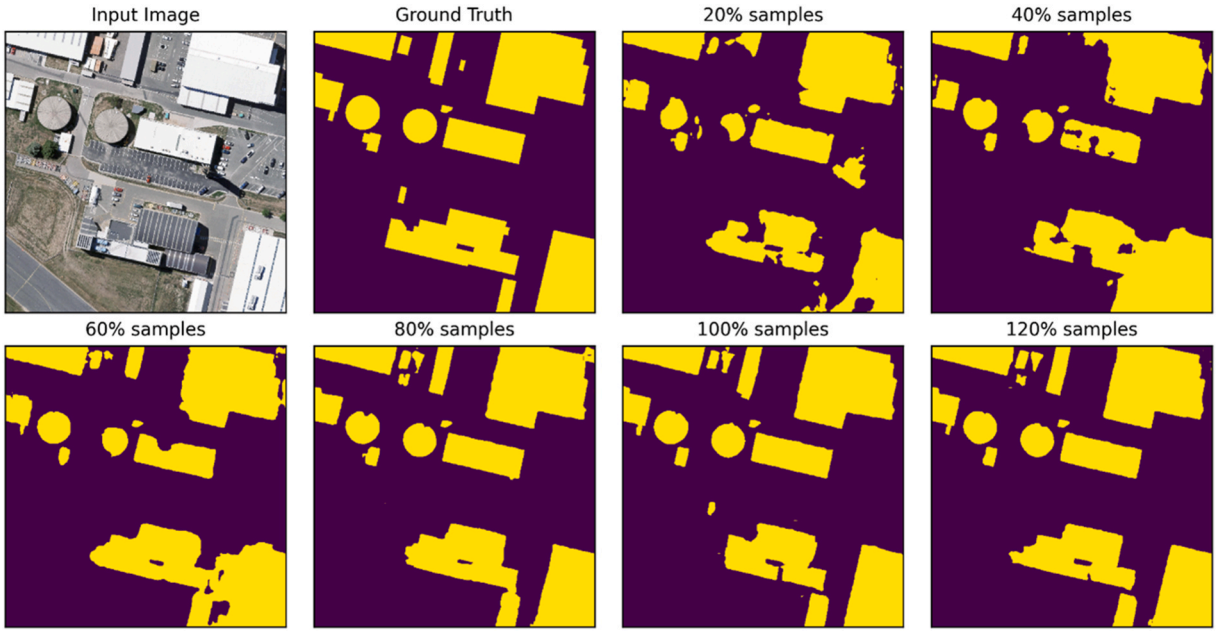

Fig. 3. Semantic segmentation results across training sample sizes using DeepLabV3+ with ResNet-152 backbone and ImageNet pretraining plus data augmentation. The first two panels show the VHR satellite image and its corresponding ground truth mask. The remaining panels illustrate the model's segmentation outputs using 20 %, 40 %, 60 %, 80 %, 100 %, and 120 % of the original training dataset. As training sample size increases, segmentation accuracy improves notably in object delineation and noise suppression, with the most refined and complete predictions observed at 100 % sample size. At 120 %, performance slightly deteriorates, likely due to overfitting or data redundancy effects, consistent with the U-shaped risk curve observed in our experiments.

Table 4
Relative improvement in validation IoU resulting from transfer learning and data augmentation across ResNet backbones under varying training data availability.

% of the original sample	Improvement in Validation accuracy by IoU (%)							
	Pre-trained ImageNet vs. Random initialisation				Pre-trained ImageNet with Data Augmentation vs. without Data Augmentation			
	ResNet-18	ResNet-50	ResNet-101	ResNet-152	ResNet-18	ResNet-50	ResNet-101	ResNet-152
20 %	0.5	1.5	3.7	6.1	3.2	2.8	2.6	3.1
40 %	4.5	6.0	6.0	7.9	0.8	0.1	1.4	1.2
60 %	3.9	4.6	4.8	6.3	1.3	0.2	0.7	1.2
80 %	3.3	3.7	3.4	5.5	0.7	0.3	0.3	0.0
100 %	2.8	3.0	3.1	3.0	0.6	0.3	0.5	0.5
120 %	2.7	2.5	2.7	5.9	1.1	0.4	0.6	0.3

meaningful improvements. ResNet-152 maintained a relatively high gain from pre-training (5.9 % IoU) even at 120 % training data (Table 4), suggesting that deeper models may continue to benefit from external knowledge under certain conditions. These improvements are consistently reflected through the F1-score (Table S2 and Fig. S2), indicating that the combined strategy of ImageNet pre-training with data augmentation yields the most robust improvements in segmentation accuracy, generalisation, and training stability.

Models trained from scratch produced sparse and inconsistent segmentation results, often failing to capture small or irregular shaped buildings (Figs. 5 and 6). Incorporating ImageNet pre-training improved delineation of structural boundaries, while the addition of data augmentation further enhanced segmentation coherence and spatial completeness. The most accurate and visually consistent results were obtained using ResNet-152 with both pre-training and augmentation, aligning with the highest quantitative improvements reported in validation IoU (Table 4). These results highlight the critical role of feature reuse and training sample diversity in supporting robust model generalisation under constrained data regimes.

3.3. Overfitting across architectures and data regimes

Detailed values for overfitting based on IoU are presented in Table 5 and Fig. 7, while corresponding results derived from the F1-score are provided in Table S3 and Fig. S3 for all ResNet backbones and training strategies. The evolution of overfitting across 50

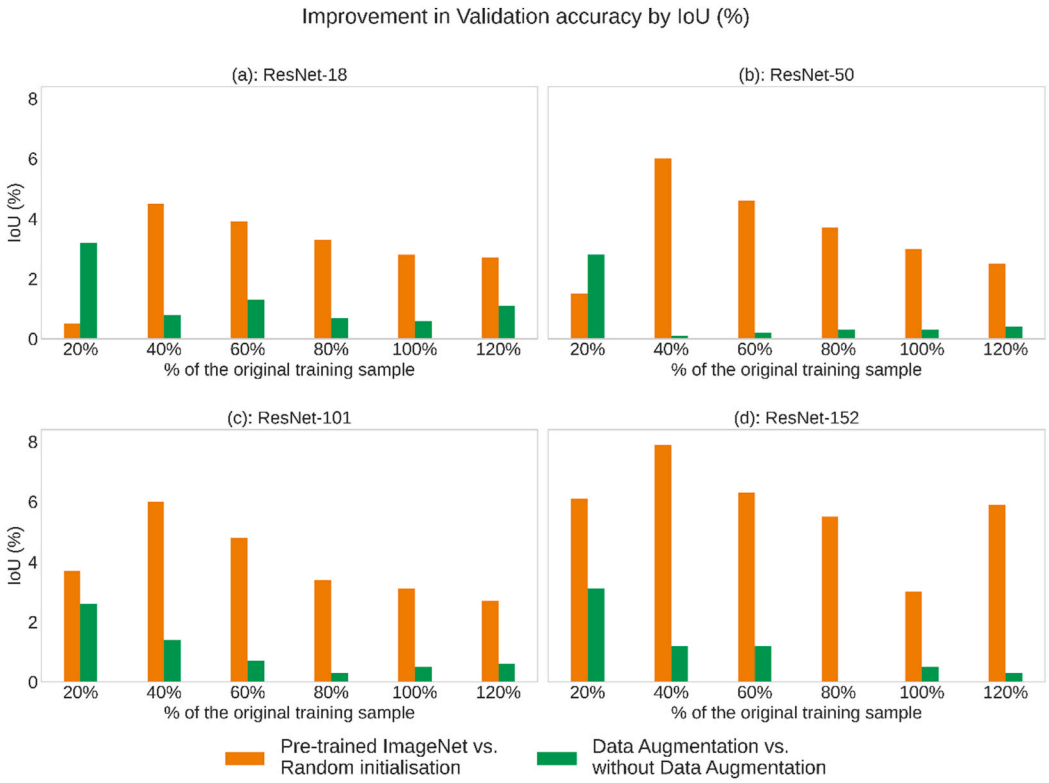

Fig. 4. Validation accuracy improvements across varying training strategies for DeepLabV3+ models with different ResNet backbones. Results compare baseline training, ImageNet-based transfer learning, and transfer learning combined with data augmentation.

training epochs for each backbone under (i) random initialisation, (ii) ImageNet pre-training, and (iii) pre-training with data augmentation was shown in Table 5 and Fig. 6. Increasing the training sample size from 20 % to 100 % consistently reduced overfittings across all models configurations. For example, with ResNet-50 using pre-training, the validation–training performance gap fell from 19.7 % at 20 % data to 6.8 % at 100 % (Table 5). Beyond this point, overfitting rose again to 10.4 % at 120 %, reflecting a U-shaped risk curve associated with excessive or redundant training samples (Fig. 7b). Across all backbones, the smallest overfitting gaps occurred at 100 % training size, identifying this regime as optimal for generalisation.

Training strategies had a significant effect on overfitting mitigation. Models initialised with ImageNet weights consistently demonstrated lower overfitting than those trained from scratch, particularly for deeper architectures. For instance, in ResNet-101, the overfitting gap dropped from 17.4 % (random initialisation) to 13.7 % (pre-trained) at 40 % training size, and from 15.2 % to 12.7 % at 60 % (Table 5; Fig. 7c). Pre-trained weights facilitated more stable optimisation by providing transferable low-level representations, reducing reliance on memorising limited training samples, especially in high-capacity networks. Data augmentation further suppressed overfitting across all sample sizes and backbones. When combined with pre-training, the validation–training gap fell below 5 % in nearly all configurations at 60 % or higher training size. At 100 % training size, overfitting was reduced to 0.5 % for ResNet-50 and 1.1 % for ResNet-152 (Table 5; Fig. 7b and d). Even under severe data constraints, augmentation yielded substantial benefits. For example, in ResNet-18 at 20 % training size, overfitting dropped from 21.2 % (pre-trained only) to 10.6 % when augmentation was applied (Fig. 7a). Consistent reductions were also observed in F1-score–based overfitting metrics (Table S3), reinforcing the stabilising effect of these regularisation strategies. These findings highlight the role of synthetic data diversity in improving generalisation across both data-scarce and data-rich regimes.

Backbone depth further influenced overfitting behaviour, with shallower architectures demonstrating greater stability under limited data conditions. At 40 % training data, ResNet-18 trained from scratch exhibited an overfitting gap of 14.3 %, which was lower than that observed for ResNet-101 (17.4 %) or ResNet-152 (18.6 %) under the same training strategy. Such differences reflect variation in model capacity, as ResNet-18 contains approximately 11 million parameters, compared with approximately 45 million for ResNet-101 and 60 million for ResNet-152, making deeper models more susceptible to overfitting when training data are scarce. However, deeper backbones proved greater benefit from advanced learning strategies. For instance, ResNet-152's overfitting gap reduced from 18.6 % (random initialisation) to 13.5 % (pre-trained), and further to 6.2 % with the addition of data augmentation at 40 % training size (Table 5; Fig. 7d). These results indicate that deeper architectures require sufficient data and stronger regularisation to achieve stable generalisation.

Fig. 5. Semantic segmentation results comparing different ResNet backbones (ResNet-18, ResNet-50, ResNet-101, ResNet-152) in DeepLabV3+, trained on 40 % of the WHU dataset using ImageNet pretraining and data augmentation. Each row presents the input image, ground truth, and segmentation predictions. Yellow represents buildings; purple represents background.

3.4. Overfitting mitigation: data augmentation and early stopping

To reduce overfitting in data-constrained environments, the impact of two widely used regularisation techniques (i.e., data augmentation and early stopping) was systematically evaluated. These strategies were tested across varying training sample sizes (20 %–120 %), model depths (ResNet-18 to ResNet-152), and training regimes (random initialisation, ImageNet pre-training, and ImageNet pre-training with data augmentation. Early stopping at 30 epochs consistently reduced overfitting compared to 50-epoch training, particularly when training data was limited. At 20 % and 40 % training sizes, all models experienced a reduction in the validation–training IoU gap when early stopping was applied (Fig. 8). For example, in ResNet-50, early stopping reduced overfitting by up to 10 % at 20 % training size, while ResNet-152 showed reductions of approximately 6–8 % across both the pre-trained and augmented configurations. These results indicates that, under limited data availability, extended training increases the risk of memorising spurious patterns, whereas earlier termination promotes more stable generalisation. Comparable improvements were observed for both IoU and F1-score metrics (Fig. 8 and Fig. S4), particularly in low-data regimes.

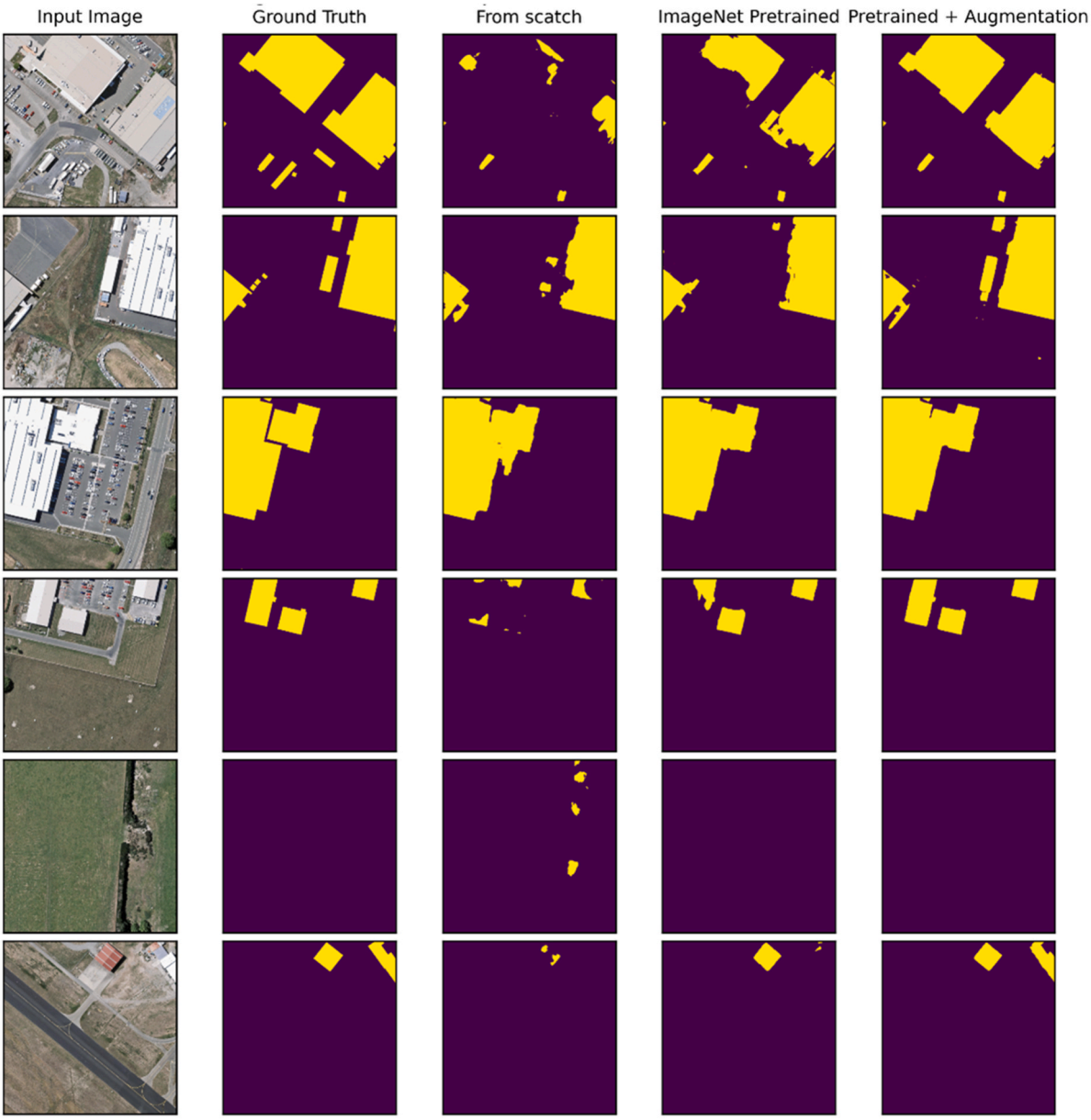

Fig. 6. Semantic segmentation results for ResNet-152 trained on 40 % of the WHU dataset under three training regimes: (i) random initialisation (from scratch), (ii) ImageNet pretraining, and (iii) ImageNet pretraining with data augmentation. Each row shows an input image, the ground truth, and corresponding segmentation outputs. Yellow indicates buildings; purple denotes background.

Data augmentation also contributed significantly to reducing overfitting across all architectures and training regimes. Across the rightmost bar groups in each plot (Fig. 5), models trained with augmentation consistently exhibited the smallest overfitting gaps, especially under early stopping conditions. In ResNet-101, for instance, the training–validation gap at 20 % training size decreased from 14.2 % (pre-trained only) to 7.3 % when augmentation and early stopping applied. Such reductions are attributed to the increased variability introduced by augmentation operations, including flipping, rotation, and intensity adjustments, which expand the effective training distribution and reduce the sensitivity to specific spatial patterns.

The combined application of data augmentation and early stopping produced the most substantial improvements. In several cases, this combination resulted in up to a 4.2 % increase in validation IoU relative to models trained without these strategies (Fig. 8). The impact was most pronounced at $\leq$40 % training sizes, where model variance and overfitting risk are highest. As training size increased beyond 60 %, performance differences between the 30- and 50-epoch training diminished, and the relative contribution of regularisation strategies decreased. At 100 % and 120 % training sizes, all configurations converged toward similarly low overfitting levels, indicating that sufficient data availability reduces dependence on strong regularisation.

Table 5
Overfitting measured as the training–validation IoU gap (%) across training sample sizes and training strategies for each ResNet backbone.

% of the original sample	Overfitting by IoU (%)											
	Random initialisation				Pre-trained imageNet				Pre-trained imageNet + DA			
	ResNet-18	ResNet-50	ResNet-101	ResNet-152	ResNet-18	ResNet-50	ResNet-101	ResNet −152	ResNet-18	ResNet-50	ResNet-101	ResNet-152
20 %	15.6	11.1	9.9	14.0	21.2	19.7	19.7	19.7	10.6	10.2	10.4	9.9
40 %	14.3	12.2	17.4	18.6	15.8	13.6	13.7	13.5	8.1	8.2	6.8	6.2
60 %	12.4	14.7	15.2	14.8	13.8	12.7	12.7	12.4	5.7	6.4	6.0	5.3
80 %	10.9	14.0	12.5	11.1	12.4	12.0	12.2	11.5	5.3	5.7	5.7	5.9
100 %	*4.7*	*8.8*	*8.6*	*8.3*	*8.1*	*6.8*	*7.3*	*7.3*	*1.5*	*0.5*	*0.9*	*1.1*
120 %	9.3	11.5	12.1	15.5	12.0	10.4	11.5	11.4	5.2	5.1	5.0	5.4

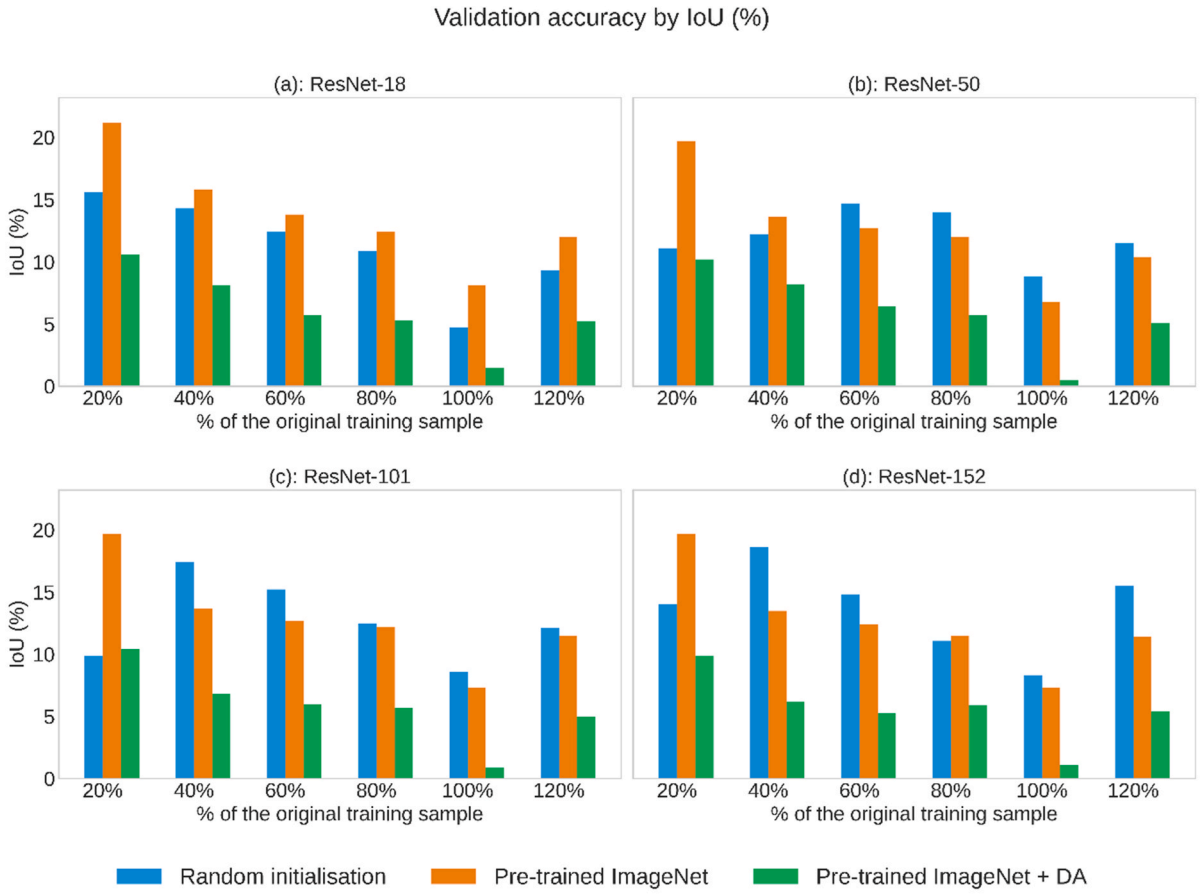

Fig. 7. Overfitting behaviour (IoU%) across ResNet backbones under varying training sample sizes (20 %–120 %) using three training strategies. Each plot compares performance under random initialisation, pre-trained ImageNet, and pre-trained ImageNet with data augmentation.

The effect of training data volume on learning dynamics is evident when comparing the 20 % and 100 % data configurations (Fig. 9). The model trained with only 20 % of the data demonstrates substantially stronger overfitting behaviour, with an overfitting gap of approximately 0.09 and a highly unstable validation loss of around 0.14, suggesting that the network has begun to memorise noise rather than capture generalisable patterns. In contrast, the model trained on the full 100 % dataset exhibits greater stability and generalisation capacity, maintaining a lower validation loss (≈0.10) and a smaller overfitting gap (≈0.05). Although both configurations show early signs of overfitting, the observed trends confirm the importance of applying early stopping at approximately epoch 30, shortly after validation loss plateaus, to improve generalisation while avoiding unnecessary computational cost.

3.5. External validation and generalisation (Japan building dataset)

To assess the robustness and generalisability of the proposed optimisation framework, an external validation experiment was conducted using the Japan Building Dataset, which represents an independent geographic and architectural domain distinct from the WHU Dataset (details provided in Supplementary Section S2 and Table S4). The validation results closely aligned with the key patterns observed in the WHU experiments, thereby confirming that the relationships among data volume, model depth, and learning strategy are not site-specific but hold across contrasting urban environments.

Validation performance, measured using both IoU and F1-score, consistently improved as the proportion of training data increased from 40 % to 100 %. Peak performance was achieved at full data utilisation, replicating the U-shaped trend identified in the WHU Dataset (Table S5–S6; Fig. S5). These findings demonstrate that larger and more diverse training sets enhance generalisation up to an optimal threshold, beyond which additional data yield diminishing returns. The effectiveness of the learning strategies also remained consistent: models initialised with pre-trained ImageNet weights outperformed those trained from scratch, and the inclusion of data augmentation provided additional gains, particularly at 40 % and 60 % training volumes (Table S7–S8; Fig. S6).

The overfitting behaviour observed in the Japan Building Dataset followed a similar pattern to that of the WHU Dataset, with the

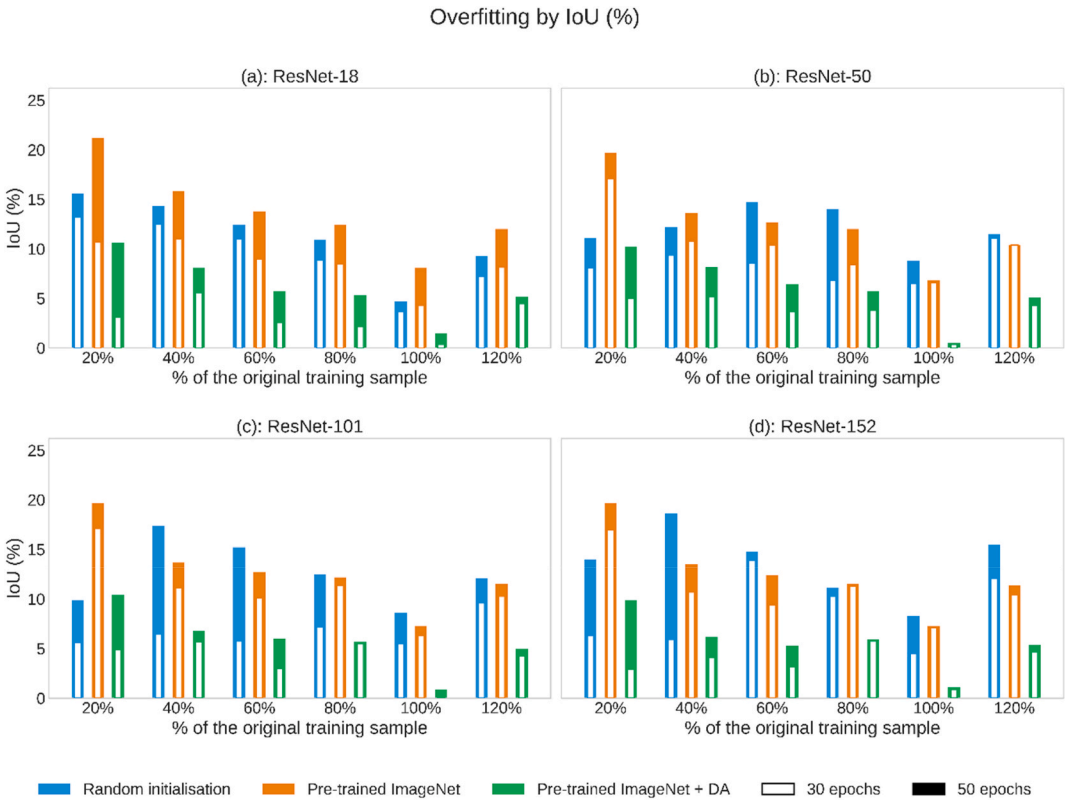

Fig. 8. Mitigation of overfitting (IoU%) through early stopping strategies applied at 30 and 50 training epochs. Performance trends are shown for DeepLabV3+ models using different backbones under three training regimes: random initialisation, ImageNet-based transfer learning, and transfer learning with data augmentation.

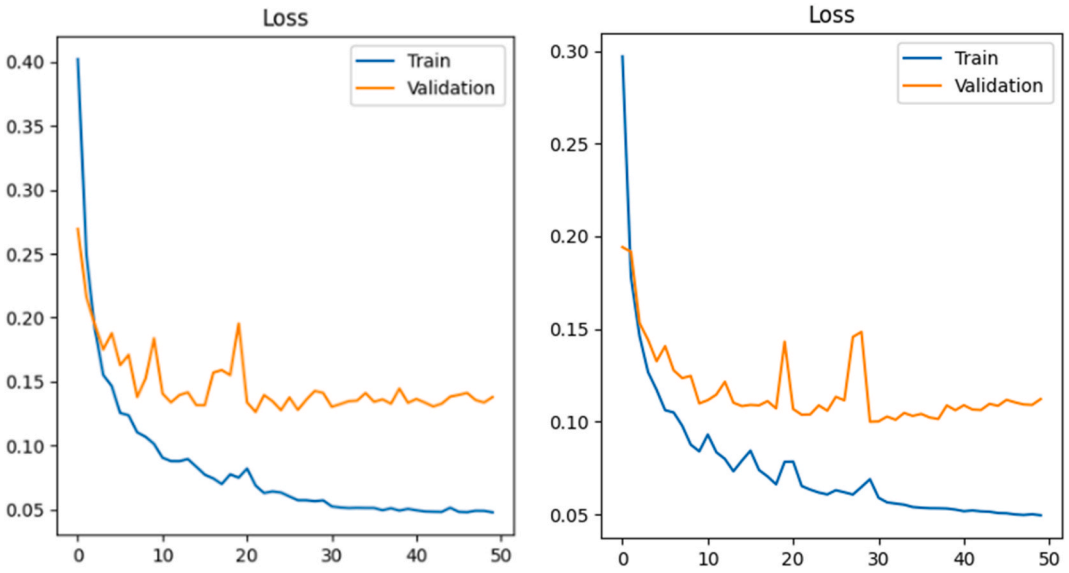

Fig. 9. Evaluating generalization loss and overfitting severity to determine optimal early stopping for 20 % and 100 % data volumes.

overfitting gap decreasing as data volume increased and reaching its minimum at 100 % training data (Table S9–S10; Fig. S7). Such consistency confirms that increased data representativeness and hybrid learning strategies enhance model stability and generalisation capacity. Across both datasets, the optimisation principles derived from the WHU experiments, relating to data thresholds, backbone

depth, and training strategy, remain valid under differing architectural styles and spatial configurations, underscoring the practical transferability of the proposed approach. Results across data regimes indicate that backbone selection is strongly dependent on training data availability. Under low-data conditions (≤40 %), shallower backbones (ResNet-18 and ResNet-50) consistently demonstrate more stable convergence and reduced overfitting compared with deeper networks. As training data volume increases (≥60 %), deeper backbones (ResNet-101 and ResNet-152) become increasingly advantageous, achieving higher accuracy without pronounced overfitting. Across all data regimes, transfer learning combined with data augmentation yields the most stable and reliable performance, supporting its selection as the preferred learning strategy under limited annotation scenarios.

3.6. Comparison with state-of-the-art methods

To assess the applicability of the proposed data-driven model selection approach relative to existing methods, peak performance was compared with representative and recent state-of-the-art (SOTA) building extraction models reported on the WHU Building Dataset (Table 6). The best-performing configuration identified in this study, DeepLabV3+ with a ResNet-152 backbone using ImageNet pre-training and data augmentation, achieved an IoU of 88.0 % and an F1-score of 93.2 %. Rather than serving as a leaderboard-style comparison, this analysis contextualises how data-informed selection of backbone depth and training strategy influences achievable performance using a standard segmentation architecture. These values place the optimised configuration within the performance range reported for widely adopted CNN-based segmentation models. The optimised model matches the IoU reported for RefineNet and DeepLabV3+ (both 88.0 %) and is only marginally lower than U-Net (88.5 % IoU), while remaining competitive with other CNN-based approaches such as SegNet, BRRNet, and DR-Net. Such performance demonstrates that aligning model complexity with data availability can yield accuracy comparable to established SOTA methods without introducing additional architectural components. Although specialised architectures such as DE-Net (90.1 % IoU, 94.8 % F1-score) and transformer-based models including SegFormer and BuildFormer report higher peak accuracy, these approaches typically rely on fixed high-capacity configurations and increased architectural complexity, often assuming abundant labelled training data.

In contrast, the present study explicitly identifies conditions under which increased backbone depth yields diminishing returns, enabling informed selection of simpler or deeper models depending on available data. Results in Table 6 indicate that competitive building extraction accuracy can be achieved through data-driven selection of backbone depth and training strategy, particularly under realistic data availability constraints. This comparison substantiates the contribution of the proposed model selection approach by demonstrating that near-SOTA performance is attainable with reduced annotation requirements and without reliance on specialised architectures, supporting scalable and interpretable deployment in very high-resolution urban mapping applications.

4. Discussion

Our study provides a comprehensive evaluation of how dataset size, model depth, and training strategy jointly influence the performance of deep learning-based building extraction under data-constrained scenarios. By systematically analysing DeepLabV3+ architectures with ResNet backbones ranging from shallow (ResNet-18) to deep (ResNet-152), across multiple training regimes and dataset proportions, the analysis reveals clear, data-dependent patterns that directly inform model selection decisions. The key findings regarding optimal dataset size, model depth, and effective regularisation strategies were further confirmed through external validation on the Japan Building Dataset (Section 3.5), demonstrating that the observed trends are not artefacts of the WHU dataset but represent generalisable selection rules applicable to VHR building extraction across diverse urban environments.

The central contribution of this study lies in explicitly linking training data availability to model selection, rather than treating backbone depth and training strategy as fixed design choices. In contrast to many peer studies that prioritise architectural novelty or peak accuracy under a single training configuration, the present analysis demonstrates how segmentation accuracy, overfitting behaviour, and convergence stability jointly determine the suitability of a given backbone under different data regimes. This data-driven selection perspective provides transparent and actionable guidance for operational deployment under realistic annotation

Table 6
Performance comparison (IoU and F1-score) between the best-performing configuration identified in this study and representative SOTA building extraction methods evaluated on the WHU Building Dataset.

Model Family	Model Architecture	IoU (%)	F1-score (%)	References
CNN-based	DeepLabV3+ (ResNet-152; ImageNet pre-training + data augmentation)	88.0	93.2	Our study
	SegNet	86.6	92.8	H. Liu et al. (2019)
	U-Net	88.5	93.9	H. Liu et al. (2019)
	RefineNet	88.0	93.6	H. Liu et al. (2019)
	DeepLabV3+	88.0	93.6	H. Liu et al. (2019)
	DE-Net	90.1	94.8	H. Liu et al. (2019)
	BRRNet	85.9	92.4	Chen et al. (2021)
	DeepLabV3+Net	85.8	92.3	Chen et al. (2021)
	DR-Net	86.0	92.5	Chen et al. (2021)
Transformer-based	SegFormer	90.1	94.9	Chen and Wu (2024)
	BuildFormer	91.4	95.5	Wang et al. (2024)

Note: SOTA results are reported values from the cited studies on the WHU building dataset; differences in data splits, preprocessing, and training protocols may affect direct comparability.

constraints.

A key finding is the non-linear relationship between training sample size and segmentation accuracy. Performance consistently improved as the training data increased from 20 % to 100 % of the WHU dataset, after which it plateaued and, in some configurations, declined slightly (Table 3 and Figs. 2 and 4). For example, validation IoU for ResNet-18 increased from 75.7 % (20 % data) to 86.6 % (100 %) but dropped to 83.1 % at 120 %. A methodological caveat applies to the 120 % scenario: the expanded dataset was intentionally constructed using samples drawn from the original test set to examine the effects of data saturation and redundancy. The observed performance degradation supports the conclusion that increasing data volume beyond the optimal threshold offers diminishing, or even negative, returns, irrespective of data origin. Such behaviour aligns with a U-shaped risk curve, where generalisation improves with additional data but may degrade when redundant or noisy samples are introduced. These findings are consistent with those reported by Zhu et al. (2020) for high-resolution segmentation tasks. The results indicate that optimal training size depends not only on quantity but also on data quality and sample distribution, a conclusion further reinforced by the Japan Dataset results (Section 3.5), where peak performance again occurred at 100 % of the available training samples.

A localised anomaly was observed (Table 3) where the ResNet-152 IoU marginally decreased from 84.4 % (60 % data) to 83.5 % (80 % data) under data augmentation. This deviation is attributed to the non-uniform characteristics of the incrementally added samples, which likely introduced a higher proportion of challenging cases (e.g., highly occluded structures or ambiguous shadow areas). Given its higher representational capacity, ResNet-152 is more susceptible to localised overfitting when exposed to such difficult features, resulting in a slight reduction in validation accuracy. This observations reinforces the broader conclusion that optimal data size is dictated by sample quality and difficulty distribution, particularly high-capacity models.

Model complexity, reflected in the depth of ResNet backbones, also played a significant role in performance outcomes. Shallower models such as ResNet-18 and ResNet-50 outperformed deeper backbones (i.e., ResNet-101, ResNet-152) under small sample regimes, particularly with random initialisation (Table 4, Fig. 4). However, as training data increased, deeper networks demonstrated superior performance, especially when combined with pretraining and augmentation. The transition point appeared to be around 60 % of the training dataset, beyond which deeper models consistently outperformed their shallower counterparts. These findings translate into explicit model selection guidance: shallower backbones (e.g., ResNet-18 and ResNet-50) are preferable under data-scarce conditions ($<$40 %), whereas deeper backbones (e.g., ResNet-101 and ResNet-152) become justified only when sufficient training data ($>$60 %) are available to support their representational capacity. These results emphasize the principle of representation efficiency, whereby model complexity should be aligned with data richness to avoid over-parameterisation and overfitting (Heryadi et al., 2020).

Transfer learning from ImageNet substantially enhanced performance across all model depths and data volumes, with gains reaching up to 8 % in some cases. The effect was most evident with deeper architectures, where pre-trained weights facilitated better feature extraction and faster convergence. Data augmentation further contributed to performance improvements, especially in low-data settings ($\leq$40 %), where it provided up to 3 % gains by synthetically enhancing training diversity. The combination of pre-training and augmentation consistently outperformed either strategy alone, suggesting a synergistic effect. These observations emphasized the importance of hybrid strategies in generalisation under limited data conditions (Balestriero et al., 2022; Shorten and Khoshgoftaar, 2019). The stability and magnitude of gains from transfer learning and data augmentation were observed to be consistent across both the WHU and Japan datasets (Supplementary Tables S7 and S8), underscoring their general applicability.

Overfitting analysis using both IoU and F1-score metrics (Table 5, Fig. 7, Supplementary Tables S3, S9, S10) further reinforced these insights. Deeper models trained without strong regularisation exhibited larger validation–training gaps, a pattern observed consistently across both datasets (i.e., WHU and Japan datasets). Pretraining and augmentation significantly reduced these gaps, particularly for ResNet-101 and ResNet-152, highlighting their role in stabilising learning dynamics. Overfitting reached its minimum at approximately 100 % training size, with a slight increase at 120 %, reinforcing the idea of an optimal training threshold. These results indicate that overfitting is governed by the interaction between model capacity, data availability, and regularisation strength, rather than model complexity alone.

Early stopping provided an additional mechanism for overfitting mitigation. Terminating training at 30 epochs, corresponding to the plateau of validation accuracy, reduced overfitting by up to 10 % in low-data scenarios (Fig. 8). The effect was most evident for ResNet-50 and ResNet-152 trained with 20–40 % of the dataset, while also reducing computational cost. Although the benefits diminished as training size increased, early stopping remains a practical and efficient strategy in data-constrained or resource-limited environments. When combined with data augmentation, early stopping yielded further improvements of up to 4.2 % in validation IoU, demonstrating complementary regularisation effects.

Compared with prior work, our study is among the few to systematically analyse the joint effects of dataset size, training strategy, and backbone complexity in VHR building extraction. While many studies employ a single configuration or limited comparisons, our results provide practical thresholds for model deployment in real-world constraints. Despite using a well-established architecture (DeepLabV3+), the optimised configuration (ResNet-152 with pre-trained ImageNet and data augmentation) achieved a peak performance of 88.0 % IoU and 93.2 % F1-score, placing it firmly within the competitive range of the current state-of-the-art (SOTA) methods in VHR segmentation (Table 6). This outcome validates the methodological focus on data efficiency and optimisation, demonstrating that competitive performance can be achieved without reliance on the most computationally demanding architectures. For practical deployment, ResNet-50 with ImageNet pretraining and data augmentation emerges as a robust default configuration under limited data availability, while deeper backbones such as ResNet-101 or ResNet-152 are recommended only when sufficient annotated data and appropriate regularisation are available. These insights are particularly relevant for applications such as disaster mapping, urban planning in low-resource regions, and UAV-based field surveys, contexts where annotated data is often scarce, and computational budgets are limited. Our findings challenge the notion that “one-size-fits-all” assumptions in deep learning and instead advocate for data-efficient modelling strategies that balance performance, stability, and scalability.

To contextualise these results within the broader literature, peak performance was compared with recent SOTA methods evaluated on the WHU dataset (Table 6). The best-performing configuration achieved 88.0 % IoU and 93.2 % F1-score, placing it within the performance range reported for established architectures such as U-Net and RefineNet. Although specialised models such as DE-Net report higher absolute accuracy, the comparison demonstrates that near-SOTA performance can be attained through informed selection of backbone depth and training strategy under realistic data and computational constraints, rather than through specialised or computationally intensive architectural innovations. The efficiency and stability achieved through the proposed hybrid strategies therefore offer a practical and scalable alternative for operational deployment in data-constrained environments.

5. Limitations and future directions

The systematic methodology developed in this study provides robust, empirically derived guidelines for optimising building extraction models under data constraints. However, several limitations define the scope of the present analysis and suggest valuable directions for future research.

A key limitation relates to the scope of generalisation across urban morphology and image characteristics. The primary experiments rely on the WHU Building Dataset, which, despite its high annotation quality, is characterised by relatively homogeneous building forms, predominantly rectangular footprints, and high-contrast urban imagery from Christchurch, New Zealand (Salman and Liu, 2019). Such structural regularity may bias the estimated performance thresholds, particularly the apparent advantage of shallower networks under low-data regimes. As a result, caution should be exercised when extrapolating these conclusions to regions with more complex, dense, or irregular urban morphologies, such as high-density Asian cities, or to imagery characterised by low contrast (Huang et al., 2016). To partially address this limitation, the Japan Building Dataset was incorporated as an independent validation domain, providing empirical evidence that the identified data–model–strategy relationships remain stable across distinct geographic and architectural contexts.

A second limitation arises from the architectural and training strategy constraints imposed to ensure experimental control. The analysis was restricted to the DeepLabV3+ architecture with ResNet backbones, selected to enable transparent isolation of backbone depth effects under varying data availability. While this constraint strengthens causal interpretation, it also limits direct extrapolation to emerging model families, including Transformer-based or hybrid CNN–ViT architectures. Future studies could extend the same controlled experimental design to these architectures to evaluate whether the observed interactions between data volume, model capacity, and learning strategy persist under fundamentally different representation mechanisms. Similarly, optimisation strategies were limited to random initialisation, transfer learning, and data augmentation. Performance gains may be further enhanced through exploration of advanced paradigms such as adversarial training, curriculum learning, dynamic learning-rate scheduling, or alternative loss functions, which were intentionally excluded to preserve experimental interpretability.

Future research would also benefit from broader geographic and contextual validation. Although the Japan Building Dataset provided an important independent test, evaluation across additional global datasets spanning diverse urban forms, sensor types, and imaging conditions would strengthen the universality of the proposed model selection principles. Beyond supervised learning, integration of semi-supervised and self-supervised approaches (Yu et al., 2023) offers a promising pathway to reduce reliance on dense manual annotation while retaining segmentation accuracy. Finally, the development of lightweight, deployable models through Transformer adaptation, neural architecture search (NAS), or model compression techniques represents a critical direction for translating data-efficient building extraction into real-time and field-ready applications.

6. Conclusions

Our study presented a systematic evaluation of DeepLabV3+ performance across varying data regimes, training strategies, and model complexities for building extraction from very high-resolution (VHR) remote sensing imagery. Through extensive experimentation with four ResNet backbones (ResNet-18, -50, −101, and −152) under six training sample sizes (20–120 % of the WHU dataset) and three learning strategies (random initialisation, transfer learning, and data augmentation), we provide comprehensive insights into how model accuracy and generalisation are shaped by key architectural and methodological choices. The evaluation explicitly links model performance to training data availability, enabling data-driven model selection rather than reliance on a single fixed configuration.

Our results reveal a non-linear relationship between training data volume and segmentation accuracy, where performance peaks at 100 % of the training data and may decline with additional samples due to redundancy and noise. Shallower models such as ResNet-18 perform best under data-scarce conditions, while deeper networks such as ResNet-101 and ResNet-152 require larger datasets to achieve optimal performance. Transfer learning with ImageNet weights significantly boosts accuracy and convergence speed, particularly in deeper models, while data augmentation proves essential for mitigating overfitting and enhancing robustness under low-data scenarios. Early stopping further emerges as a lightweight yet effective strategy for improving generalisation while reducing training cost.

These outcomes were further validated through an external experiment using the Japan Building Dataset, confirming that the optimised configurations identified with WHU data are transferable across distinct urban contexts. The optimised DeepLabV3+ model with a ResNet-152 backbone, ImageNet pre-training, and data augmentation achieved 88.0 % IoU and 93.2 % F1-score, positioning it within the competitive range of state-of-the-art methods. Such performance substantiates the central premise of the study, demonstrating that systematic optimisation guided by data availability and training strategy can achieve near-SOTA accuracy without reliance on specialised or computationally intensive architectural innovations.

The study highlights the importance of aligning model complexity with data availability and incorporating targeted regularisation strategies. Rather than defaulting to the deepest architectures or the largest possible datasets, practitioners are encouraged to balance computational cost, data accessibility, and model stability. Based on the empirical evidence presented, ResNet-50 combined with transfer learning and data augmentation emerges as a robust and resource-efficient default configuration for data-constrained applications, while deeper backbones are recommended only when sufficient annotated data and appropriate regularisation are available. By offering empirically grounded model selection guidance, this research contributes to the development of scalable and data-efficient approaches for VHR building extraction and urban analytics.

CRediT authorship contribution statement

Dung T. Pham: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Thuong V. Tran:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Xuan Zhu:** Writing – review & editing, Writing – original draft, Validation, Supervision. **Hung N. Pham:** Validation, Formal analysis, Data curation.

Funding

This research is funded by the Ministry of Education and Training of Vietnam under Grant Number B2024-MDA-09 to D.T.P and H. N.P.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Dung T. Pham reports financial support was provided by Hanoi University of Mining and Geology. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rsase.2026.101876.

Data availability

Data will be made available on request.

References

Abriha, D., Szabó, S., 2023. Strategies in Training Deep Learning Models to Extract Building from Multisource Images with Small Training Sample Sizes, 16, pp. 1707–1724. %J I.J. of D.E.

Ait El Asri, S., Negabi, Ismail, El, Adib Samir, Raissouni, N., 2023. Enhancing building extraction from remote sensing images through UNet and transfer learning. Int. J. Comput. Appl. 45, 413–419. https://doi.org/10.1080/1206212X.2023.2219117.

Al-Ani, S., Guo, H., Fyfe, S., Long, Z., Donnaz, S., Kim, Y., 2025. Effect of training sample size, image resolution and epochs on filamentous and floc-forming bacteria classification using machine learning. J. Environ. Manag. 379, 124803. https://doi.org/10.1016/j.jenvman.2025.124803.

Alnagashi, F., Rahim, N.A., Shukor, S.A.A., Hamid, M.H.A.%J.A.M.C.I., 2024. Mitigating Overfitting in Extreme Learning Machine Classifier Through Dropout Regularization, 13, pp. 26–35.

Alomar, K., Aysel, H.I., Cai, X., 2023. Data Augmentation in Classification and Segmentation: a Survey and New Strategies, 9, p. 46.

Ayala, C., Sesma, R., Aranda, C., Galar, M.%J.R.S., 2021. A Deep Learning Approach to an Enhanced Building Footprint and Road Detection in high-resolution Satellite Imagery, 13, p. 3135.

Bai, G., Chai, Z., Ling, C., Wang, S., Lu, J., Zhang, N., Shi, T., Yu, Z., Zhu, M., Zhang, Y., Yang, C., Cheng, Y., Zhao, L., 2024. Beyond efficiency: a systematic survey of resource-efficient large language models. arXiv:2401.00625 4, 1–70.

Bakirman, T., Komurcu, I., Sertel, E., 2022. Comparative Analysis of Deep Learning Based Building Extraction Methods with the New VHR Istanbul Dataset, 202, 117346. %J E.S. with A.

Balestriero, R., Bottou, L., LeCun, Y., 2022. The Effects of Regularization and Data Augmentation are Class Dependent, 35, pp. 37878–37891. %J A. in N.I.P.S.

Cai, S., Shu, Y., Chen, G., Ooi, B.C., Wang, W., Zhang, M., 2019. Effective and efficient dropout for deep convolutional neural networks. arXiv preprint arXiv: 03392.

Cha, Y.-J., Ali, R., Lewis, J., Büyüköztürk, O., 2024. Deep learning-based structural health monitoring. Autom. Constr. 161, 105328.

Chen, J., Wu, Q., 2024. Improved DeepLabv3+ connected augmented reality technology for building target extraction in urban environmental design. Int. J. Inf. Commun. Technol. 24, 54–73.

Chen, M., Wu, J., Liu, L., Zhao, W., Tian, F., Shen, Q., Zhao, B., Du, R., 2021. DR-Net: an improved network for building extraction from high resolution remote sensing image. Remote Sens. 13, 294. https://doi.org/10.3390/rs13020294.

Chen, S., Ogawa, Y., Zhao, C., Sekimoto, Y., 2023. Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach. ISPRS J. Photogrammetry Remote Sens. 195, 129–152. https://doi.org/10.1016/j.isprsjprs.2022.11.006.

Cherti, M., Jitsev, J., 2022. Effect of pre-training Scale on Intra- and inter-domain full and few-shot transfer learning for natural and medical X-Ray chest images. In: 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1–9. https://doi.org/10.1109/IJCNN55064.2022.9892393.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: a large-scale hierarchical image database. Presented at the 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. Ieee.
Du, Shouji, Du, Shihong, Liu, B., Zhang, X., 2021. Incorporating DeepLabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images. Int. J. Digit. Earth 14, 357–378.
Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.%J.I., journal of computer vision, 2010. The pascal visual object classes (voc) challenge 88, 303–338.
Fernandez-Moral, E., Martins, R., Wolf, D., Rives, P., 2018. A new metric for evaluating semantic segmentation: leveraging global and contour accuracy. Presented at the 2018 IEEE Intelligent Vehicles Symposium (Iv), pp. 1051–1056. IEEE.
Gao, Y., Ruan, Y., Fang, C., Yin, S., 2020. Deep Learning and Transfer Learning Models of Energy Consumption Forecasting for a Building with Poor Information Data, 223, 110156. %J E., Buildings.
Ghaffarian, S., Kerle, N., Filatova, T., 2018. Remote sensing-based proxies for urban disaster risk management and resilience: A review. Remote Sens. 10, 1760.
Guo, M., Liu, H., Xu, Y., Huang, Y.%J.R.S., 2020. Building Extraction Based on U-Net with an Attention Block and Multiple Losses, 12, p. 1400.
Hay, G.J., Castilla, G., 2006. Object-based image analysis: strengths, weaknesses, opportunities and threats (SWOT). Presented at the Proc. 1st Int. Conf. OBIA, pp. 4–5. Citeseer.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.
He, K., Zhang, X., Ren, S., Sun, J.%J.I., 2015. Transactions on pattern analysis, machine intelligence. Spatial pyramid pooling in deep convolutional networks for visual recognition 37, 1904–1916.
He, M., Zhang, J., He, Y., Zuo, X., Gao, Z.%J.R.S., 2024. Annotated dataset for training cloud segmentation neural networks using high-resolution satellite. Remote Sensing Imagery 16, 3682.
Heryadi, Y., Irwansyah, E., Miranda, E., Soeparno, H., Hashimoto, K., 2020. The effect of resnet model as feature extractor network to performance of DeepLabV3 model for semantic satellite image segmentation. Presented at the 2020 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS), pp. 74–77. IEEE.
%J I.J. of S.T. in A.E.O., Remote Sensing Huang, X., Yuan, W., Li, J., Zhang, L., 2016. A new building extraction postprocessing framework for high-spatial-resolution remote-sensing imagery, 10, 654–668.
Iannelli, G.C., Lisini, G., Dell'Acqua, F., Feitosa, R.Q., da Costa, G.A.O.P., Gamba, P., 2014. Urban area extent extraction in spaceborne HR and VHR data using multi-resolution features. Sensors 14, 18337–18352.
Ji, S., Wei, S., Lu, M., 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery. data set 57, 574–586. %J I.T. on geoscience, remote sensing.
Koldasbayeva, D., Tregubova, P., Gasanov, M., Zaytsev, A., Petrovskaia, A., Burnaev, E.%J.N.C., 2024. Challenges in data-driven Geospatial Modeling for Environmental Research and Practice, 15, 10700.
Kumar, M., Bhardwaj, A., 2020. Building extraction from very high resolution stereo satellite images using OBIA and topographic information. Environmental Sciences Proceedings 5, 1. https://doi.org/10.3390/IECG2020-08908.
Li, S., Song, W., Fang, L., Chen, Y., Ghamisi, P., Benediktsson, J.A., 2019. Deep learning for hyperspectral image classification: an overview. IEEE Trans. Geosci. Rem. Sens. 57, 6690–6709.
Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: common objects in context. Presented at the Computer vision–ECCV 2014: 13Th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, pp. 740–755.
Liu, H., Luo, J., Huang, B., Hu, X., Sun, Y., Yang, Y., Xu, N., Zhou, N., 2019. DE-Net: deep encoding network for building extraction from high-resolution remote sensing imagery. Remote Sens. 11, 2380. https://doi.org/10.3390/rs11202380.
Liu, W., Yue, A., Shi, W., Ji, J., Deng, R., 2019. An automatic extraction architecture of urban green space based on DeepLabv3plus semantic segmentation model. Presented at the 2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC). IEEE, pp. 311–315.
Liu, Y., Fan, B., Wang, L., Bai, J., Xiang, S., Pan, C.%J.I., 2018. Journal of photogrammetry, remote sensing. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network 145, 78–95.
Luo, L., Li, P., Yan, X., 2021. Deep learning-based building extraction from remote sensing images: a comprehensive review. Energies 14, 7982. https://doi.org/10.3390/en14237982.
Neupane, B., Aryal, J., Rajabifard, A., 2025. Fine-Tuning-Based Transfer Learning for Building Extraction from Off-Nadir Remote Sensing Images 17. %J R.S.
Panigrahi, S., Nanda, A., Swarnkar, T., 2020. A survey on transfer learning. Intelligent and Cloud Computing: Proceedings of ICICC 2019 1, 781–789. Springer.
Patel, G.M., Shukla, S., 2025. Automated extraction of building footprints from high-resolution Cartosat-2 imagery for urban planning. Indian J. Sci. Technol. 18, 1000–1008.
Prakash, P.S., Soni, J., Bharath, H.A., 2022. Building extraction from remote sensing images using deep learning and transfer learning. Presented at the IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, pp. 3079–3082. IEEE.
Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S., 2019. Generalized intersection over union: a metric and a loss for bounding box regression. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 658–666.
Salman, S., Liu, X., 2019. Overfitting mechanism and avoidance in deep neural networks. arXiv preprint arXiv: 06566.
Santos, C.F.G.D., Papa, J.P., 2022. Avoiding overfitting: a survey on regularization methods for Convolutional Neural Networks. ACM Comput. Surv. 54, 1–25.
Shan, Z., Liu, Y., Zhou, L., Yan, C., Wang, H., Xie, X., 2025. ROS-SAM: High-Quality interactive segmentation for remote sensing moving object. https://doi.org/10.48550/arXiv.2503.12006.
Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6, 60. https://doi.org/10.1186/s40537-019-0197-0.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.%J.T., 2014. Dropout: a Simple Way to Prevent Neural Networks from Overfitting, 15, pp. 1929–1958 journal of machine learning research.
Vincent M, J., P, V., 2024. Extraction of building footprint using MASK-RCNN for high resolution aerial imagery. Environ. Res. Commun. 6, 075015. https://doi.org/10.1088/2515-7620/ad5b3d.
Wang, C., Du, P., Wu, H., Li, J., Zhao, C., Zhu, H.%J.C., 2021. A Cucumber Leaf Disease Severity Classification Method Based on the Fusion of DeepLabV3+ and U-Net, 189, 106373 electronics in agriculture.
Wang, Y., Yang, L., Liu, X., Yan, P., 2024. An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabv3+. Sci. Rep. 14, 9716. https://doi.org/10.1038/s41598-024-60375-1.
Wangiyana, S., Samczyński, P., Gromek, A., 2022. Data Augmentation for Building Footprint Segmentation in SAR Images: an Empirical Study, 14, p. 2012.
Wei, Y.-C., Chen, W.-L., Tuanmu, M.-N., Lu, S.-S., Shiao, M.-T., 2024. Advanced montane bird monitoring using self-supervised learning and transformer on passive acoustic data. Ecol. Inform. 84, 102927. https://doi.org/10.1016/j.ecoinf.2024.102927.
Yan, Y., Tan, Z., Su, N., Zhao, C., 2017. Building extraction based on an optimized stacked sparse autoencoder of structure and training samples using LIDAR DSM and optical images. Sensors (Basel) 17. https://doi.org/10.3390/s17091957.
%J I.J. of S.T. in A.E.O., Remote Sensing Yang, H.L., Yuan, J., Lunga, D., Laverdiere, M., Rose, A., Bhaduri, B., 2018. Building extraction at scale using convolutional neural network: mapping of the United States, 11, 2600–2614.
Ye, Z., Yang, K., Lin, Y., Guo, S., Sun, Y., Chen, X., Lai, R., Zhang, H., 2023. A comparison between Pixel-based deep learning and object-based image analysis (OBIA) for individual detection of cabbage plants based on UAV visible-light images. Comput. Electron. Agric. 209, 107822. https://doi.org/10.1016/j.compag.2023.107822.
Yu, A., Quan, Y., Yu, R., Guo, W., Wang, X., Hong, D., Zhang, H., Chen, J., Hu, Q., He, P.%J.R.S., 2023. Deep learning methods for semantic segmentation in remote sensing with small data. A survey 15, 4987.

Zhang, L., Wu, J., Fan, Y., Gao, H., Shao, Y.%J.S., 2020. An Efficient Building Extraction Method from High Spatial Resolution Remote Sensing Images Based on Improved Mask, 20. R-CNN, p. 1465.
Zhao, X., Li, S., Sun, Z., Ma, H., Kong, X., Xu, Y.%J.R.S.L., 2025. Detection of Building Shadows in high-resolution Remote Sensing Images by Using Improved DeepLabV3+, 16, pp. 290–301.
Zhu, Q., Li, Z., Zhang, Y., Guan, Q.%J.R.S., 2020. Building Extraction from High Spatial Resolution Remote Sensing Images via multiscale-aware and segmentation-prior Conditional Random Fields, 12, p. 3983.