International Journal of Engineering and Geosciences
https://dergipark.org.tr/en/pub/ijeg
e-ISSN 2548-0960

The Role of Ensemble Deep Learning for Building Extraction from VHR Imagery

Nuran Aslantas [1], Tolga Bakirman [*2], Mahmut Oğuz Selbesoğlu [3], Bülent Bayram[2]

[1] *Yildiz Technical University, Geomatics Engineering Department, Türkiye, nuranaslantas25@gmail.com*
[2] *Yildiz Technical University, Geomatics Engineering Department, Türkiye, bakirman@yildiz.edu.tr*
[3] *Istanbul Technical University, Geomatics Engineering Department, Türkiye, selbesoglu@itu.edu.tr*
[4] *Yildiz Technical University, Geomatics Engineering Department, Türkiye, bayram@yildiz.edu.tr*

Cite this study: Aslantas, N., Bakirman, T., Selbesoğlu, M.O. & Bayram, B. (2024). The Role of Ensemble Deep Learning for Building Extraction from VHR Imagery. International Journal of Engineering and Geosciences, 10(3), 352-363.

https://doi.org/10.26833/ijeg.1587798

Keywords
Remote Sensing
Deep Learning
Building Extraction
VHR Imagery
Ensemble Model

Research Article

Received:
Revised:
Accepted:
Published:

Abstract

In modern geographical applications, the demand for up-to-date and accurate building maps is increasing, driven by essential needs in sustainable urban planning, sprawl monitoring, natural hazard mitigation, crisis management, smart city initiatives, and the establishment of climate-resilient urban environments. The unregulated growth in urbanization and settlement patterns poses multifaceted challenges, including ecological imbalances, loss of arable land, and increasing risk of drought. Leveraging recent technologies in remote sensing and artificial intelligence, particularly in the fields of very high-resolution satellite imagery and aerial photography, presents promising solutions for rapidly acquiring precise building maps. This research aims to investigate the efficiency of an ensemble deep learning framework comprising DeepLabV3+, UNet++, Pix2pix, Feature Pyramid Network, and Pyramid Scene Parsing Network architectures for the semantic segmentation of buildings. By employing the Wuhan University Aerial Building Dataset, characterized by a spatial resolution of 0.3 meters, as the training and testing dataset, the study assesses the performance of the proposed ensemble model. The findings reveal notable accuracies, with intersection over union metrics reaching 90.22% for DeepLabV3+, 91.01% for UNet++, 83.50% for Pix2pix, 88.90% for FPN, 88.20% for PSPNet, and finally at 91.06% for the ensemble model. These results reveal the potential of integrating diverse deep learning architectures to enhance the precision of building semantic segmentation.

1. Introduction

The exponential growth of the global population, surpassing 8 billion individuals, has precipitated a significant shift towards urbanization, with more than half of humanity now concentrated in urban environments—a stark departure from the mere 2% recorded in 1800 [1]. While urban centers serve as hubs for governance, economic vitality, cultural exchange, and collective learning [2], the unbounded expansion of urban populations has engendered unsustainable urbanization practices, adversely impacting species diversity [3], wildlife habitats [4], climate stability [5], and critical ecosystem services [6]. In light of these challenges, maintaining accurate and up-to-date information regarding urban infrastructure, particularly buildings, assumes vital importance for effective urban planning, infrastructure development, and timely emergency response to natural disasters [7-13] Furthermore, optimizing building density and distribution emerges as a crucial strategy in fostering climate-resilient and smart urban development [14,15].

Remote sensing technologies offer a robust solution to rapidly acquire high-resolution data across extensive geographic areas, for the mapping and monitoring of urban structures [16-21]. However, the extraction of buildings from remote sensing imagery remains a challenge due to complex backgrounds, diverse roof types, and occlusions [22]. To address this challenge, researchers have explored various methodologies, including physical rule-based approaches, object-oriented methods, supervised classification, machine learning, and deep learning techniques. Recent investigations into building extraction methods using very high-resolution optical remote sensing imagery have been extensively reviewed by [23].

Deep learning methodologies have found widespread application across various domains, including building extraction, subsequent to their triumph in the ImageNet challenge [24]. Notably, researchers have introduced and implemented deep learning architectures tailored specifically for building extraction applications [25,26]. Ensemble deep learning strategies have been adopted to integrate multiple models, thereby addressing issues related to inaccurate pixel classification [27,28]. [29] presents one of the pioneering instances of ensemble deep learning applied to building extraction using very high resolution (VHR) remote sensing imagery, wherein semantic segmentation is integrated with edge detection. Their approach involves integrating the Holistically-Nested Edge Detection network into both SegNet and Fully Convolutional Network (FCN), followed by ensembling SegNet and two FCN variants, subsequently validated with VHR imagery and digital surface models from ISPRS Potsdam and Vaihingen datasets. [30] devised an ensemble model comprising feature extractors from various architectures including AlexNet, VGG-Net, GoogLeNet, and SqueezeNet, customized in terms of filters, depth, and window size. Their ensemble model is proposed for village building identification utilizing Google and Bing imagery with 1.2-meter spatial resolution. [31] trained four distinct U-Net architectures with pansharpened WorldView-3 imagery and a fusion of satellite imagery and GIS map images, employing rescaled and sliced images for training. The resultant probability maps are aggregated through ensemble methods to facilitate the final prediction. [27] leveraged multiple U-Net architectures with varying parameter configurations, encompassing general parameters, increased building loss ratio, geometric transformations, and stricter building selection criteria. Their approach involves integrating prediction outcomes from different U-Net architectures trained on SuperView-1 imagery with a 50 cm resolution, utilizing a voting method for integration. [32] introduced a stacking ensemble model comprising U-Net, SegNet, and FCN-8s architectures. Their methodology includes conditional random field post-processing for each segmentation result, alongside sparse autoencoder utilization to encode multilayer features and align the primary encoder output with real data. Experiments were conducted using WorldView and QuickBird series imagery to validate their approach. [33] developed an ensemble network consisting of U-Net and SegNet named Seg-Unet for building extraction from Massachusetts building dataset. [34] utilized ensemble techniques for multi-source data for building segmentation using various open-access datasets in the literature.

Despite these advancements in the literature, the utilization of ensemble deep learning for building extraction remains limited, with existing applications predominantly relying on conventional models such as U-Net, SegNet, and FCN, thereby overlooking recent state-of-the-art architectures.

Addressing this gap, this study aims to develop an ensemble deep learning model comprising DeepLabV3+, UNet++, conditional generative adversarial network (cGAN), Feature Pyramid Network (FPN), and Pyramid Scene Parsing Network (PSPNet) architectures for building extraction from very high-resolution imagery. The primary motivation behind this endeavor lies in harnessing the collective power of diverse deep learning architectures to enhance the accuracy and robustness of building extraction algorithms. By integrating recent state-of-the-art models into an ensemble framework, this research seeks to push the boundaries of existing methodologies and pave the way for more effective and efficient urban monitoring and management strategies.

2. Method

In this study, we utilized the open-access Wuhan University (WHU) building dataset, as detailed in the study by [35] . The dataset comprises both aerial and satellite images, with a specific focus on the aerial building dataset. This subset encompasses 8,189 natural image tiles, each measuring 512 × 512 pixels, and exhibits a spatial resolution of 0.30 meters, down sampled from the original 0.075-meter resolution data. Notably, the dataset encompasses 187,000 individual buildings located in Christchurch, New Zealand. The labels for these buildings were precisely generated through manual editing of the original vector data obtained from the land information service of New Zealand. Importantly, the dataset offers comprehensive coverage of diverse urban landscapes, encompassing rural, residential, cultural, and industrial areas within the city. For visual reference, sample image tiles and their corresponding labels from the dataset is provided in Figure 1.

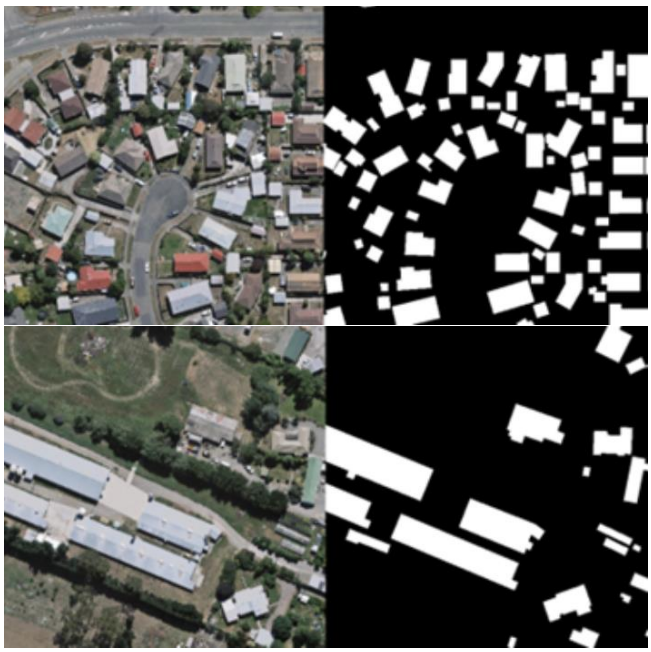


Figure 1. Sample 512 x 512 pixels size image tiles and corresponding building labels from the WHU buildings

During this investigation, we partitioned the WHU dataset into distinct subsets designated for training, validation, and testing purposes. The training set comprises 4,736 image tiles, encompassing 130,500

building labels, while the validation set consists of 1,036 image tiles containing 14,500 building labels. Additionally, the test set comprises 2,416 image tiles, containing a total of 42,000 building labels.

We employed five distinct deep learning architectures, specifically DeepLabv3+, UNet++, Pix2Pix, Feature Pyramid Network (FPN), and Pyramid Scene Parsing Network (PSPNet). Notably, [25] conducted a comprehensive assessment of DeepLabv3+, U-Net, UNet++, FPN, and PSPNet architectures for building segmentation, utilizing a proprietary dataset from Istanbul. Their findings demonstrated promising performance across various datasets, thus motivating our adoption of DeepLabv3+, UNet++, FPN, and PSPNet architectures. Furthermore, instead of the U-Net architecture, we opted for the Pix2Pix model owing to its demonstrated efficacy in prior studies, leveraging its conditional generative adversarial network (cGAN) structure.

DeepLab represents a sophisticated semantic segmentation model developed under Google, constituting an open-source framework. The evolution of the DeepLab architecture has been notable, transitioning from its initial version in 2014 [36] to the more recent DeepLabV3+ iteration introduced in 2018 [37]. Distinguished by its fusion of encoder-decoder network principles and spatial pyramid pooling techniques, DeepLabV3+ capitalizes on a modified Xception architecture, strategically employing atrous convolutions to improve issues associated with low-resolution features. Furthermore, the model incorporates atrous spatial pyramid pooling, facilitating the comprehensive representation of objects across multiple scales. Depth wise separable convolution, integral to both Atrous Spatial Pyramid Pooling (ASPP) and decoder modules, enhances network efficiency and robustness. ASPP comprises multiple layers of parallel convolutions with varying dilation rates, each generating distinct feature maps, subsequently associated into a cohesive final feature map. This multiplicity of parallel convolution layers enables the model to perceive objects and contextualize images across diverse scales. To refine segmentation outcomes and delineate object boundaries more precisely, a dedicated decoder module is introduced. Through this intricate design, DeepLabV3+ exhibits a high capacity for accurate image segmentation, particularly in the delineation of fine boundaries.

The UNet++ architecture, proposed by [38], constitutes a deep-trained encoder-decoder network characterized by a complex arrangement of nested and dense skip connections linking the encoder and decoder subnetworks. Unlike its predecessor, the original U-Net, UNet++ distinguishes itself through three key modifications: redesigned skip pathways, incorporation of dense skip connections, and implementation of deep supervision mechanisms. In the redesigned skip pathways of the UNet++ architecture, the output from the preceding convolution layer within the same dense block is fused with the upsampled output of the corresponding lower dense block. Dense skip connections are strategically integrated into the skip pathways between the encoder and decoder components, drawing inspiration from DenseNet [39] to enhance segmentation accuracy and optimize gradient flow. Notably, dense scatter connections facilitate the aggregation of all preceding feature maps, ensuring their propagation to the current node through dense convolution blocks along each jump link, thereby generating multiple full-resolution feature maps. Deep supervision is a pivotal design aspect of UNet++, involving the creation of multi-resolution segmentation maps. This deep supervision mechanism serves as an effective regularization technique, fostering classification accuracy and facilitating feature learning in scenarios characterized by limited training data and relatively shallow network architectures.

The concept of utilizing Adversarial Networks to generate images was originally introduced by [40]. In the training of Generative Adversarial Networks (GANs), two networks engage in a simultaneous competition: a generator network produces synthetic data intended to closely match the distribution of the training data, while a discriminator network aims to distinguish between real and counterfeit data. In the present study, we have employed Pix2Pix [41] for conditional Generative Adversarial Network (cGAN) implementation.

Pix2Pix is designed to learn mapping from an input image to generate a corresponding output image, employing a defined loss function during training [41]. Its adaptability allows it to be applied to a variety of problem domains, as it is not constrained to specific applications. The architecture of Pix2Pix encompasses a U-Net based design for the generator component, while a PatchGAN classifier is utilized on the discriminator side. To address challenges related to generating high-resolution output and mitigating bottlenecks associated with progressively down sampling inputs, skip connections are incorporated between each layer within the generator. On the discriminator side, PatchGAN operates to distinguish the realness of images based on N x N patches, traversing the entire image to compute a final output through averaging responses. This approach ensures the modelling of high-frequency details by leveraging the structural information present in local image patches.

The Feature Pyramid Network (FPN), as introduced by [42], represents a fully convolutional neural network framework tailored for semantic segmentation tasks. FPN integrates feature pyramids, a critical component utilized for the detection of objects across varying scales. These pyramids facilitate the systematic scanning of the model across both spatial locations and pyramid levels, thereby enabling the detection of objects across a broad spectrum of scales, resulting in the generation of multi-scale feature maps. Central to the functionality of FPN is the fusion of information derived from both lower and higher-level features to inform prediction processes. This is achieved through a top-down architectural design supplemented by lateral connections for the generation of high-level multi-scale feature maps. Importantly, each pyramid level produced within the FPN architecture maintains consistent channel dimensions. While FPN is

primarily devised for object identification tasks, its adaptability extends to encompassing segmentation tasks as well.

PSPNet proposed by [43] represents a widely adopted neural network architecture, commonly utilizing ResNet backbones to construct the initial feature map. This model capitalizes on the global context of input images through the incorporation of a pyramid pooling module. Within the encoder subnetwork, convolutional neural networks are employed to extract feature maps [44]. Subsequently, a pyramid decomposition module is applied to gather diverse subregion representations, followed by super-sampling and merging layers for the generation of a comprehensive final feature representation that encapsulates both local and global contextual information. The pyramid pooling module operates across four distinct pyramid scales, facilitating the aggregation of features. Subsequently, the feature maps are organized into multiple groups and upsampled to their original dimensions, before being combined with the original feature maps. This strategic integration ensures the preservation and fusion of both low-level and high-level features, thereby enriching the overall representation with a blend of local and global contextual characteristics. This sophisticated architecture enables PSPNet to effectively capture complex spatial relationships and semantic information within images, thereby facilitating accurate segmentation outcomes across a diverse array of tasks and datasets.

Ensemble methods deviate from conventional learning approaches by constructing a collection of models and fusing their outputs, thereby leveraging the complementary strengths of individual models to enhance predictive performance. Unlike traditional methods that aims to formulate a singular model from training data, ensemble methods accommodate a spectrum of model complexities, ranging from weak learners capable of surpassing random guessing to potent models adept at delivering precise predictions. The significance of ensemble methods dates to seminal works by [45] and [46], which laid the foundation for their widespread adoption since the 1990s. Typically, the construction of an ensemble unfolds in two primary stages: first, the development of constituent core models, and subsequently, their fusion into a cohesive ensemble.

In this study, we employed the majority voting method to craft our ensemble model. Despite its apparent simplicity, the majority voting method has exhibited efficacy across diverse applications within the literature, as evidenced by studies conducted by [47], [48], and [49]. Within the majority voting framework, each model contributes a class label vote for every pixel, with the final prediction being determined by the class that garners more than half of the collective votes. Detailed definition of the majority voting processes can be found in the seminal work by [50], which offers insights into the operational mechanics underlying this straightforward yet effective ensemble method. Through the adoption of the majority voting approach, our study aims to capitalize on its demonstrated efficacy and simplicity, thereby yielding robust and reliable ensemble predictions in the context of semantic segmentation tasks.

3. Results

In this study, the WHU building dataset served as the main data source for training, validation, and testing procedures, conducted within the Python environment utilizing the PyTorch library. The training iterations of DeepLabV3+, UNet++, PSPNet, and FPN architectures were executed on a computing system equipped with a Nvidia Tesla P100 PCIE 16 GB GPU. Conversely, training sessions for the Pix2Pix architecture were conducted on a separate computing setup featuring an NVIDIA GeForce GTX1080 Ti GPU with 11 GB RAM, alongside an Intel® CoreTM i7-8700K 3.70GHz processor. A schematic representation of the study's workflow is provided in Figure 2. Detailed specifications of the hyperparameters utilized for each deep learning architecture are delineated in Table 1. For Pix2pix, Binary Cross Entropy was chosen as it aligns with its adversarial training framework, ensuring stable optimization and consistency with GAN-based image translation practices.

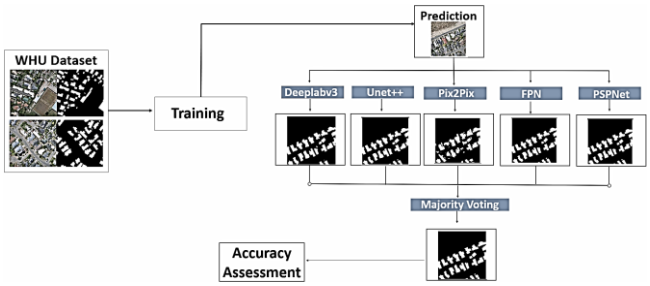


Figure 2. General workflow of the study

Table 1. Hyperparameters used for training process of each DL architecture.

	DeepLabv3+	Unet++	PSPNet	FPN	Pix2Pix
Train	4736				4736
Validation	1036				-
Test	2416				2416
Epochs	30				30
Loss Function	Dice Loss				Binary Cross Entropy
Activation Function	Sigmoid				Sigmoid
Optimizer	Adam				Adam
Batch Size	4				1

Furthermore, an invetigation was conducted on the efficiency of various encoders, specifically EfficientNet-b6, SE-ResNeXt101, and InceptionResNetv2, within the context of DeepLabV3+, UNet++, PSPNet, and FPN architectures, guided by their documented efficiencies in prior literature as highlighted by [25]. Comprehensive accuracy metrics related to each architecture as well as the ensemble network are detailed in Table 2. The EfficientNet encoder exhibited superior performance across the DeepLabv3+, UNet++, and FPN architectures. Conversely, the SE-ResNext encoder demonstrated a

slight advantage over the EfficientNet encoder within the PSPNet architecture.

Table 2. Accuracy metrics derived from the evaluation with the test dataset.

(%)	Encoder	Acc.	IoU	Prec.	Rec.	F1
DeepLabv3+	EfficientNet	98.86	90.22	95.44	94.25	94.85
Unet++	EfficientNet	98.96	91.01	95.60	94.98	95.29
PSPNet	SE-ResNeXt	98.62	88.20	94.85	92.62	93.72
FPN	EfficientNet	98.87	90.27	95.30	94.47	94.88
Pix2Pix	-	98.01	83.50	90.97	91.03	91.00
Emsemble	-	98.96	91.06	95.94	94.70	95.32

The comprehensive evaluation of accuracy metrics reveals that the proposed ensemble network surpassed all individual architecture employed in terms of all accuracy metrics, except with a marginal difference in recall. Specifically, the ensemble network achieved accuracy, IoU, precision, recall, and F1-score values of 98.96%, 91.06%, 95.94%, 94.70%, and 95.32%, respectively.

Notably, while UNet++ achieved the highest scores when considered individually, both UNet++ and the ensemble network yielded identical accuracy values. However, the recall value of UNet++ surpassed that of the proposed ensemble network. Example prediction results are presented in Figure 3 for reference.

Visual examination of the prediction outputs indicates a degree of similarity among the results obtained from DeepLabv3+, UNet++, and FPN. However, PSPNet exhibits the tendency to generate false-positive pixels, particularly manifesting as connections between adjacent buildings, as illustrated in Figure 3(a5). Moreover, Pix2Pix manifests small false-positive noisy patches, often attributed to vehicles.

DeepLabv3+ and UNet++ excel in preserving the shapes of buildings, particularly with respect to their edges and minor protrusions. Furthermore, these architectures demonstrate proficiency in accurately extracting small structures such as sheds or garages, as depicted in Figure 3b.

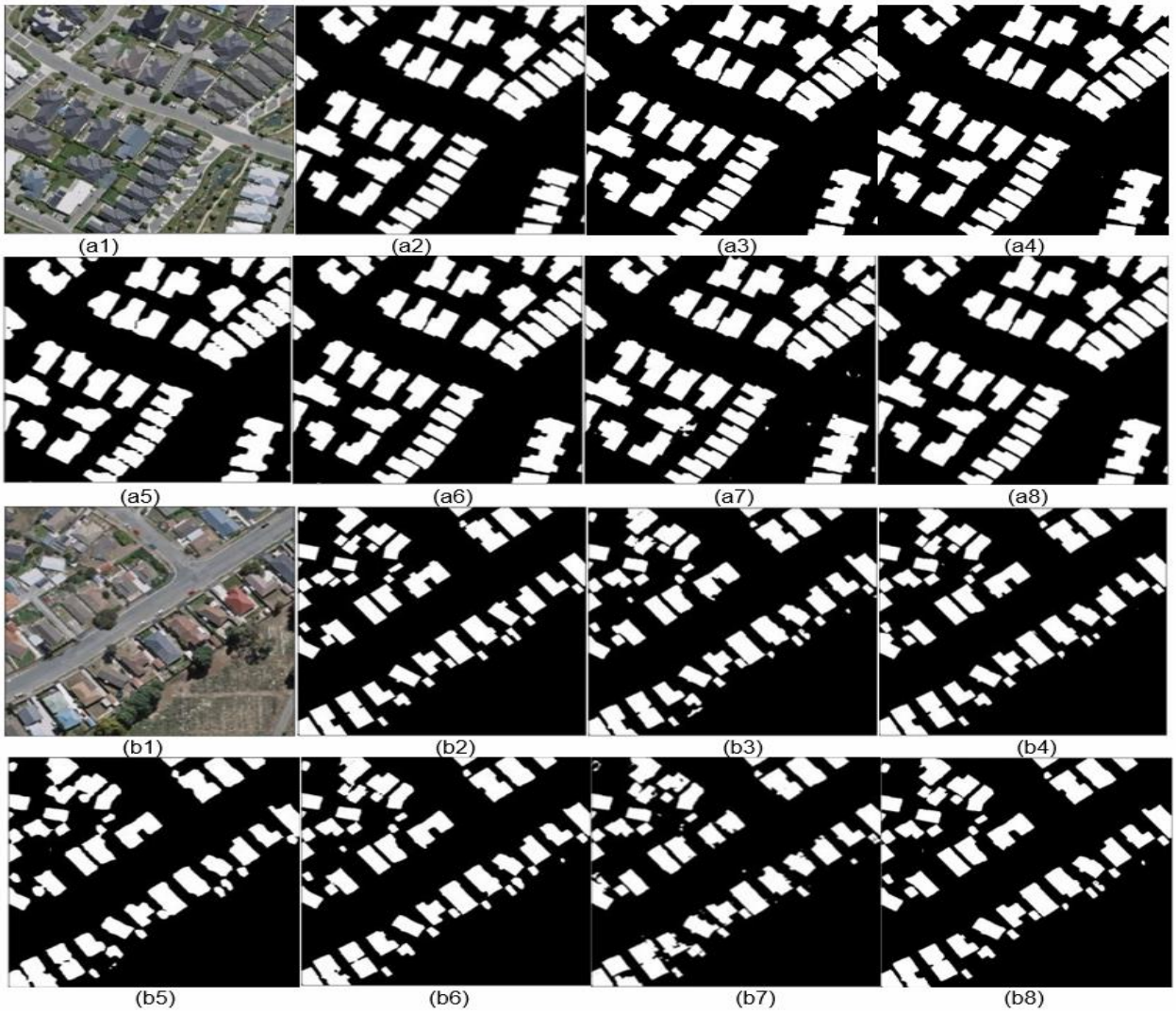
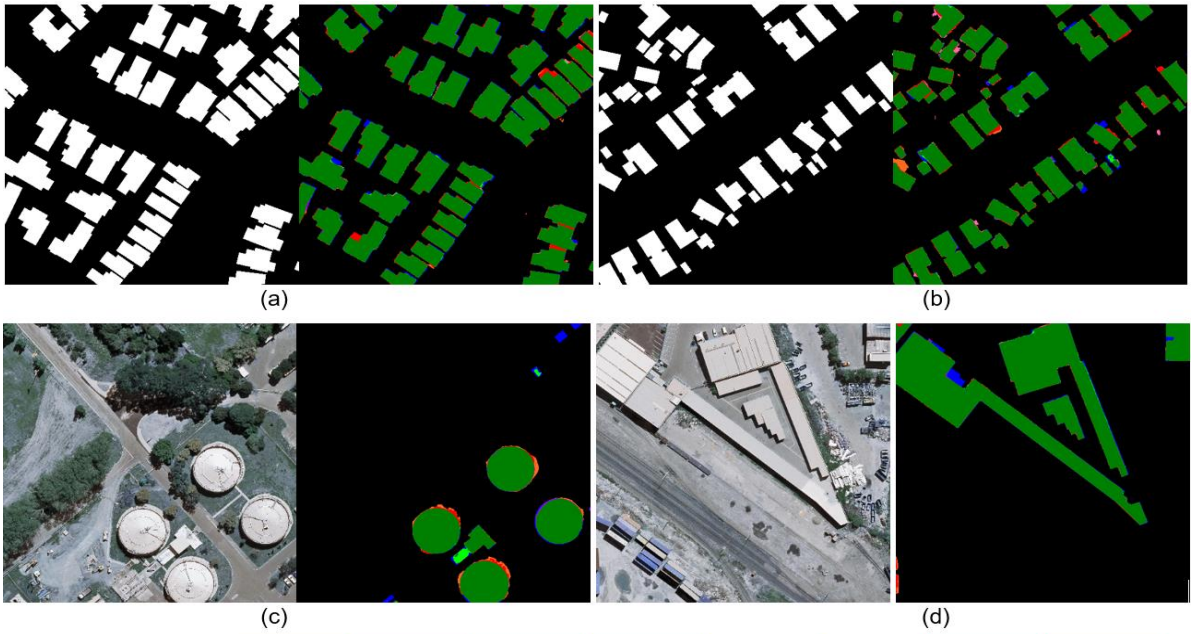


Figure 3. Prediction outcomes derived from the evaluation with the test dataset. The figure displays: (1) the original test image, (2) the corresponding ground truth, and the segmented predictions generated by (3) DeepLabv3+, (4) UNet++, (5) PSPNet, (6) FPN, (7) Pix2Pix, and (8) the Ensemble Network.



Color Code								
Unet++	TP	TP	FN	FN	TN	FP	TN	FP
Ensemble	TP	FN	TP	FN	TN	TN	FP	FP

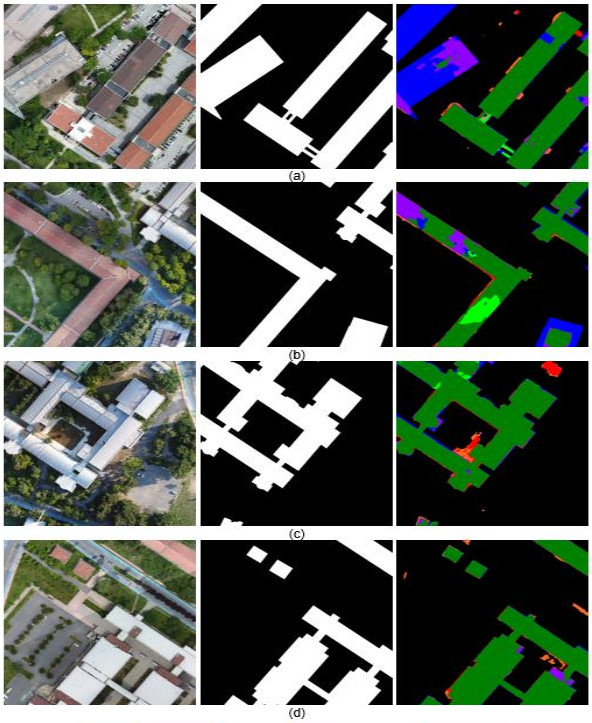
Figure 4. Samples illustrating the comparison between the ensemble network and UNet++ (Left: Input Image, Right: Prediction comparison). (a), (b), (c) and (d) are examples for different types of buildings. (a) & (b): Residential buildings, (c): Round buildings, (d): Industrial buildings

Given UNet++'s superior individual performance, comprehensive comparisons were conducted between UNet++ and the ensemble network, as depicted in Figure 4. These comparisons juxtapose the prediction results with the ground truth, facilitating a detailed assessment. Notably, the ensemble network demonstrates competency in accurately delineating irregularly shaped buildings, including those with rounded or pointed features, as evidenced in Figure 4(c) and 4(d). Despite slight distortions along the edges, the network successfully captures the general structural attributes of rounded buildings (Figure 4c), while also preserving sharp edges, even in instances of narrow or pointed building segments (Figure 4d). These findings underscore the efficacy of the proposed ensemble network in extracting buildings with well-preserved edges and minimal false-positive identifications.

4. Discussion

To further explore and test generalization the capabilities of the ensemble network, additional tests were conducted using images independent from the WHU dataset. In the initial test, predictions were executed on an ortho-photo obtained from an Unmanned Aerial Vehicle (UAV) over Istanbul. The raw images were captured using a DJI Phantom 4 equipped with a true-color camera, resulting in an orthophoto with a ground sample distance of 3 cm. However, similar to the approach adopted for the WHU building dataset, the orthophoto image was down sampled to a 30 cm resolution. This downsampling process, while necessary for consistency with the training data, likely resulted in the loss of fine-grained details that are critical for accurate building extraction in HR imagery. Comparative analysis of the results revealed that while the ensemble network exhibited a low incidence of false positives, it struggled to accurately extract significant portions of buildings, resulting in a notable number of false negatives (Figure 5). UNet++ demonstrated a similar performance, except with a higher occurrence of false positives (Figure 5c and 5d). This discrepancy between false positives and false negatives is reflected in the precision and recall values for each image, as detailed in Table 3.

The challenges observed in this test can be attributed to several factors specific to high-resolution imagery. First, the loss of detail due to downsampling may have blurred or eliminated distinguishing features, particularly for buildings with concrete roofs, which were often entirely missed (Figure 5a). Second, the increased complexity of high-resolution imagery, with its finer textures and patterns, likely contributed to the misclassification of pathways as buildings, especially when they exhibited similar spectral or textural characteristics to nearby roofs (Figure 5c). Additionally, cloud shadows significantly impacted segmentation performance, as evidenced in Figure 5b. This sensitivity to environmental factors is a known challenge in high-resolution imagery, where variations in lighting and shadows are more pronounced and can adversely affect model performance. These findings highlight the limitations of the current model when applied to high-resolution data and underscore the need for future improvements.



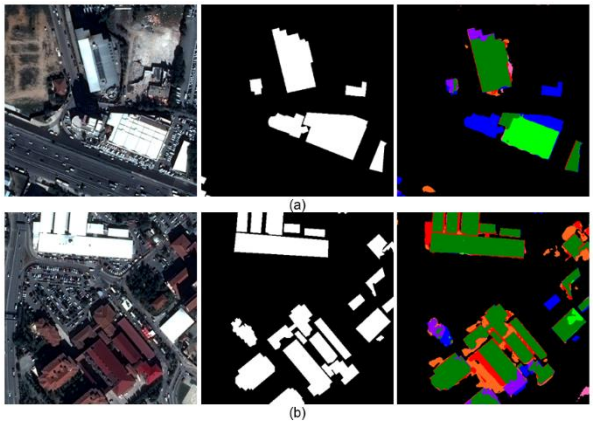
Color Code								
Unet++	TP	TP	FN	FN	TN	FP	TN	FP
Ensemble	TP	FN	TP	FN	TN	TN	FP	FP

Figure 5. Prediction examples for UAV images from Istanbul. Left Column: Test Image, Middle Column: Ground Truth, Right Column: Prediction Comparison.

Table 3. Accuracy metrics for UAV Images from Istanbul using the ensemble network (first line) and UNet++ (second line).

First Line: The Ensemble Network
Second Line: UNet++

	Accuracy	IoU	Precision	Recall	F1
Figure 5(a)	83.33 85.88	57.56 64.61	98.50 96.43	58.07 66.20	73.07 78.50
Figure 5(b)	94.15 94.55	76.36 78.35	97.89 96.00	77.63 81.00	86.59 87.86
Figure 5(c)	96.99 96.69	89.84 89.04	94.83 92.83	94.47 95.62	94.65 94.20
Figure 5(d)	98.65 98.51	94.60 94.19	98.38 95.50	96.09 98.57	97.23 97.01



Color Code								
Unet++	TP	TP	FN	FN	TN	FP	TN	FP
Ensemble	TP	FN	TP	FN	TN	TN	FP	FP

Figure 6. Prediction examples for Pléiades imagery from Istanbul. Left Column: Test Image, Middle Column: Ground Truth, Right Column: Prediction Comparison.

Table 4. Accuracy metrics for Pléiades imagery from Istanbul using the ensemble network (first line) and UNet++ (second line). The best results from the Istanbul dataset [25] are also given for reference.

First Line: The Ensemble Network
Second Line: UNet++

	Accuracy	IoU	Precision	Recall	F1
Figure 6(a)	95.07 91.64	66.14 43.30	94.98 90.23	68.54 45.43	79.62 60.44
Figure 6(b)	92.82 90.09	74.56 69.29	83.73 73.35	87.20 92.60	85.43 81.86
Istanbul Dataset	95.52	93.80	95.97	97.53	93.80

In the presented study, an ensemble network comprising DeepLabv3+, UNet++, PSPNet, FPN, and Pix2Pix architectures was developed. The proposed network demonstrates commendable performance on the trained WHU dataset, yielding precise outcomes when compared to recent literature. A comprehensive overview of the accuracy results obtained with the WHU aerial dataset is presented in Table 5.

Table 5. Accuracy metrics derived from evaluations conducted with the WHU aerial dataset as reported in recent literature. *Baseline denotes the baseline method. A "-" indicates that the accuracy metric value is not provided in the corresponding paper. (Arranged chronologically).

Architecture	IoU (%)	Precision (%)	Recall (%)	F1-Score (%)
Siamese U-Net* [35]	88.40	93.80	93.90	-
RFA-UNet [51]	90.02	-	-	94.75
EU-Net [52]	90.56	94.98	**95.10 (9th)**	95.04
SRI-Net [53]	89.09	95.21	93.28	94.23
DE-Net [54]	90.12	95.00	94.60	94.80
ESFNet [55]	85.34	-	-	-
SR-FCN [56]	88.90	94.40	93.90	-
HFSA-Unet [57]	90.72	95.09	**95.18 (8th)**	95.13
ENRU-Net [58]	90.77	-	-	95.16
ARC-Net [59]	**91.80 (6th)**	**96.40 (4th)**	**95.10 (9th)**	**95.70 (5th)**
MSCRF [60]	**91.99 (5th)**	95.07	**96.47 (4th)**	-
PISANet [61]	87.97	94.20	92.94	93.55
BuildingNAS [62]	86.95	-	-	-
EANet [63]	**93.33 (3rd)**	**98.67 (1st)**	**96.42 (5th)**	**97.52 (2nd)**
Attention-Based [64]	90.29	94.97	**94.81 (11th)**	94.90
DR-Net [65]	88.30	-	-	93.80
Self-Attention U-Net [66]	89.39	93.25	**95.56 (6th)**	94.40
SST [67]	89.01	-	-	94.13
BE Network [68]	86.15	91.76	93.37	91.97
DLEBFP [69]	85.10	92.60	91.40	-
CT-Unet [70]	91.00	94.95	94.02	-
csAG-HRNet [71]	-	**98.42 (2nd)**	**98.70 (2nd)**	**98.55 (1st)**
RU-Net [22]	**94.61 (1st)**	**97.48 (3rd)**	**96.98 (3rd)**	**97.23 (3rd)**
GCCINet [72]	78.88	89.09	87.31	88.19
SCGF ConvNets [73]	90.90	**96.00 (5th)**	94.60	95.30
CA-BASNet [74]	**93.43 (4th)**	90.13	**98.79 (1st)**	91.35
MSL-Net [75]	90.40	95.10	**94.80 (12th)**	95.00
AEUNet++ [76]	**91.08 (7th)**	95.23	**95.43 (7th)**	**95.33 (6th)**
STEB-UNet [77]	**93.89 (2nd)**	-	-	**96.85 (4th)**
Boundary DCNN [78]	89.97	94.99	94.45	94.72
ASGASN [79]	89.40	93.80	**95.10 (9th)**	94.40
CFENet [80]	87.22	-	-	92.01
Unet++ w/ EfficientNet (Ours)	**91.01**	**95.60**	**94.98 (10th)**	**95.29**
Ensemble Network (Ours)	**91.06 (8th)**	**95.94 (6th)**	**94.70 (12th)**	**95.32 (7th)**

Based on the literature review, our proposed ensemble model ranked 8th, 6th, 13th, and 7th for IoU, precision, recall, and F1-score, respectively, among the studies utilizing the WHU aerial building dataset. The average deviation between the highest accuracy metric value reported in the literature and our findings is approximately 3% for all accuracy metrics. This variation may be attributed to the utilization of the same architecture in the ensemble model without structural modifications. Nevertheless, the outcomes remain promising and effective, particularly when considering the examination of 32 studies published between 2019 and 2022.

5. Conclusion

The main objective of this study is to generate precise building maps from VHR aerial imagery, which constitute a fundamental element of digital twin frameworks for urban environments. To achieve this goal, we have devised an ensemble model leveraging contemporary deep learning architectures, utilizing the publicly available WHU aerial building dataset. Our findings demonstrate the efficiency of ensemble models in segmenting buildings from VHR imagery. While our proposed ensemble network and UNet++ exhibited comparable performance based on accuracy metrics, comparative analyses revealed that our ensemble network effectively mitigates false positives, a challenge encountered with UNet++. Hence, it can be inferred that the integration of other architectures contributes to the ensemble model's ability to reduce false positives.

Furthermore, our ensemble model underwent testing across various scales and with different sensors. While our model, trained on aerial data, exhibited satisfactory performance when applied to UAV imagery, similar results could not be replicated with VHR optical satellite imagery, as the predictions were deemed insufficient for further analysis. Consequently, plans are underway to conduct experiments using 30cm-resolution VHR satellite imagery to address this limitation.

The implications of our study extend to various domains including population growth management, disaster preparedness, environmental monitoring, and

sustainable resource utilization. Moreover, our methodology holds promises for facilitating digital twin generation, offering developments that prioritize both environmental sustainability and digital transformation. The automated analysis of existing settlement patterns and building mapping using deep learning techniques presents an efficient, rapid, reliable, and labor-saving approach that aligns with the objectives of digitalization, data reuse, and citizen science.

In conclusion, while the results obtained in this study are promising, our future studies aim to refine our approach by developing a novel deep learning ensemble model encompassing a broader array of architectures. This approach will enable us to achieve superior results and facilitate comparative assessments against various existing ensemble models in forthcoming studies.

Acknowledgement

The authors declare the utilization of AI-assisted technologies to enhance the readability and linguistic quality of the manuscript.

Author contributions

Nuran Aslantas: Conceptualization, Data curation, Methodology, Software, Validation **Tolga Bakirman:** Data curation, Writing-Original draft preparation, Software, Validation, Visualization **Mahmut Oğuz Selbesoğlu:** Investigation, Writing-Reviewing and Editing **Bülent Bayram:** Conceptualization, Investigation, Methodology, Supervision, Writing-Reviewing and Editing

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Cohen, J. E. (2003). Human population: the next half century. *Science*, *302*(5648), 1172-1175.
2. Rees, W. E. (1999). The built environment and the ecosphere: a global perspective. *Building Research & Information*, *27*(4-5), 206-220.
3. McKinney, M. L. (2008). Effects of urbanization on species richness: a review of plants and animals. *Urban ecosystems*, *11*, 161-176.
4. Marzluff, J. M. (2001). Worldwide urbanization and its effects on birds. *Avian ecology conservation in an urbanizing world*, 19-47.
5. Grimmond, S. (2007). Urbanization and global environmental change: local effects of urban warming. *The Geographical Journal*, *173*(1), 83-88.
6. Chen, W., & Chi, G. (2022). Urbanization and ecosystem services: The multi-scale spatial spillover effects and spatial variations. *Land Use Policy*, *114*, 105964.
7. Abdollahi, A., Pradhan, B., Gite, S., & Alamri, A. (2020). Building footprint extraction from high resolution aerial images using generative adversarial network (GAN) architecture. *IEEE Access*, *8*, 209517-209527.
8. Mousa, Y. A., Helmholz, P., Belton, D., & Bulatov, D. (2019). Building detection and regularisation using DSM and imagery information. *The Photogrammetric Record*, *34*(165), 85-107.
9. Habib, W., Mahmood, S., Noor, S., Saleem, A., Siraj, M., & Ahmad, H. (2023). A post earthquake damage assessment using GIS in district Mirpur, Pakistan. *Advanced GIS*, *3*(2), 53-58.
10. Pehlivan, H. (2021). The Analysis Methodology of Robotic Total Station Data for Determination of Structural Displacements. *Advanced Geomatics*, *1*(1), 1-7.
11. Pala, İ., & Algancı, U. Investigating the performance of super-resolved remote sensing images on coastline segmentation with deep learning based methods. *International Journal of Engineering Geosciences*, *10*(1), 93-106.
12. Varul, Y. E., Adıyaman, H., Bakırman, T., Bayram, B., Alkan, E., Karaca, S. Z., & Topaloğlu, R. H. (2023). Preserving human privacy in real estate listing applications by deep learning methods. *Mersin Photogrammetry Journal*, *5*(1), 10-17.
13. Bovkır, R. (2024). İstanbul'da kentsel yeşil altyapı için çatı tarımı potansiyelinin CBS tabanlı karar analizi ile değerlendirilmesi. *Geomatik*, *10*(1), 45-58.
14. Jabareen, Y. (2013). Planning the resilient city: Concepts and strategies for coping with climate change and environmental risk. *Cities*, *31*, 220-229.
15. Liu, R., Wang, M., Hou, G., Wu, W., Zhao, C., & Ge, Q. (2023). The classification of airborne LiDAR building point clouds based on multi-scale and multi-level cloth simulation. *The Photogrammetric Record*, *38*(182), 118-136.
16. Akbulut, Z., Özdemir, S., Acar, H., Dihkan, M., & Karslı, F. (2018). Automatic extraction of building boundaries from high resolution images with active contour segmentation. *International Journal of Engineering Geosciences*, *3*(1), 36-42.
17. Tabakoğlu, C. (2024). A Review: Detection types and systems in remote sensing. *Advanced GIS*, *4*(2), 100-105.
18. Sariturk, B., Bayram, B., Duran, Z., & Seker, D. Z. (2020). Feature extraction from satellite images using segnet and fully convolutional networks (FCN). *International Journal of Engineering Geosciences*, *5*(3), 138-143.
19. Şenol, H. İ., Kaya, Y., Yiğit, A. Y., & Yakar, M. (2024). Extraction and geospatial analysis of the Hersek Lagoon shoreline with Sentinel-2 satellite data. Survey Review, 56(397), 367-382.
20. Ünyılmaz, C. N., Kulavuz, B., Bakırman, T., & Bayram, B. (2024). Deep Learning Based Panchromatic2RGB Image Generation from VHR Images. *Mersin Photogrammetry Journal*, *6*(2), 87-92.
21. Bayram, B., Kılıç, B., Özoğlu, F., Erdem, F., Bakirman, T., Sivri, S., Delen, A. (2020). A deep learning integrated mobile application for historic landmark recognition: A case study of Istanbul. *Mersin Photogrammetry Journal*, *2*(2), 38-50.

22. Wang, H., & Miao, F. (2022). Building extraction from remote sensing images using deep residual U-Net. *European Journal of Remote Sensing*, *55*(1), 71-85.
23. Li, J., Huang, X., Tu, L., Zhang, T., & Wang, L. (2022a). A review of building detection from very high resolution optical remote sensing images. *GIScience Remote Sensing*, *59*(1), 1199-1225.
24. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.
25. Bakirman, T., Komurcu, I., & Sertel, E. (2022). Comparative analysis of deep learning based building extraction methods with the new VHR Istanbul dataset. *Expert Systems with Applications*, *202*, 117346.
26. Şengül, G. S., & Sertel, E. (2024). Automatic Building Extraction From VHR Remote Sensing Images Using Geoai Methods. IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium,
27. Chen, D.-Y., Peng, L., Li, W.-C., & Wang, Y.-D. (2021a). Building extraction and number statistics in WUI areas based on UNet structure and ensemble learning. *Remote Sensing*, *13*(6), 1172.
28. Erdem, F., Bayram, B., Bakirman, T., Bayrak, O. C., & Akpinar, B. (2021). An ensemble deep learning based shoreline segmentation approach (WaterNet) from Landsat 8 OLI images. *Advances in Space Research*, *67*(3), 964-974.
29. Marmanis, D., Schindler, K., Wegner, J. D., Galliani, S., Datcu, M., & Stilla, U. (2018). Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry Remote Sensing*, *135*, 158-172.
30. Guo, Z., Chen, Q., Wu, G., Xu, Y., Shibasaki, R., & Shao, X. (2017). Village building identification based on ensemble convolutional neural networks. *Sensors*, *17*(11), 2487.
31. Li, W., He, C., Fang, J., Zheng, J., Fu, H., & Yu, L. (2019). Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data. *Remote Sensing*, *11*(4), 403.
32. Kaya, Y., Şenol, H. İ., Yiğit, A. Y., & Yakar, M. (2023). Car detection from very high-resolution UAV images using deep learning algorithms. Photogrammetric Engineering & Remote Sensing, 89(2), 117-123.
33. Abdollahi, A., Pradhan, B., & Alamri, A. (2022). An ensemble architecture of deep convolutional Segnet and Unet networks for building semantic segmentation from high-resolution aerial images. *Geocarto International*, *37*(12), 3355-3370.
34. Wang, S., Zang, Q., Zhao, D., Fang, C., Quan, D., Wan, Y., Jiao, L. (2023). Select, purify, and exchange: A multisource unsupervised domain adaptation method for building extraction. *IEEE Transactions on Neural Networks Learning Systems*.
35. Ji, S., Wei, S., & Lu, M. (2019a). Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on geoscience remote sensing*, *57*(1), 574-586.
36. Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2015). Semantic image segmentation with deep convolutional nets and fully connected crfs. International conference on learning representations, San Diego, United States.
37. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV),
38. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4,
39. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition,
40. Akçay, Ö., Erenoğlu, R. C., & Avşar, E. Ö. (2017). The effect of jpeg compression in close range photogrammetry. International Journal of Engineering and Geosciences, 2(1), 35-40. https://doi.org/10.26833/ijeg.287308.
41. Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition,
42. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition,
43. Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. Proceedings of the IEEE conference on computer vision and pattern recognition,
44. Bekçi, R. N., Zorlu, Ö., & Menekşe, E. (2022). Regression analysis and use of artificial neural networks in housing valuation forecasting: case example of Güvenevler neighbourhood in Mersin. *Advanced GIS*, *2*(1), 24-32.
45. Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE transactions on pattern analysis machine intelligence*, *12*(10), 993-1001.
46. Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, *5*, 197-227.
47. Pleşoianu, A.-I., Stupariu, M.-S., Şandric, I., Pătru-Stupariu, I., & Drăguţ, L. (2020). Individual tree-crown detection and species classification in very high-resolution remote sensing imagery using a deep learning ensemble model. *Remote Sensing*, *12*(15), 2426.
48. Ben Jabra, M., Koubaa, A., Benjdira, B., Ammar, A., & Hamam, H. (2021). COVID-19 diagnosis in chest X-

rays using deep learning and majority voting. *Applied Sciences*, *11*(6), 2884.
49. Tandel, G. S., Tiwari, A., & Kakde, O. G. (2021). Performance optimisation of deep learning models using majority voting algorithm for brain tumour classification. *Computers in Biology Medicine*, *135*, 104564.
50. Lam, L., & Suen, C. Y. (1995). Optimal combinations of pattern classifiers. *Pattern Recognition Letters*, *16*(9), 945-954.
51. Ye, Z., Fu, Y., Gan, M., Deng, J., Comber, A., & Wang, K. (2019). Building extraction from very high resolution aerial imagery using joint attention deep neural network. *Remote Sensing*, *11*(24), 2970.
52. Kang, W., Xiang, Y., Wang, F., & You, H. (2019). EU-Net: An efficient fully convolutional network for building extraction from optical remote sensing images. *Remote Sensing*, *11*(23), 2813.
53. Liu, P., Liu, X., Liu, M., Shi, Q., Yang, J., Xu, X., & Zhang, Y. (2019b). Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network. *Remote Sensing*, *11*(7), 830.
54. Liu, H., Luo, J., Huang, B., Hu, X., Sun, Y., Yang, Y.,…Zhou, N. (2019a). DE-Net: Deep encoding network for building extraction from high-resolution remote sensing imagery. *Remote Sensing*, *11*(20), 2380.
55. Lin, J., Jing, W., Song, H., & Chen, G. (2019). ESFNet: Efficient network for building extraction from high-resolution aerial images. *IEEE Access*, *7*, 54285-54294.
56. Ji, S., Wei, S., & Lu, M. (2019b). A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery. International Journal of Remote Sensing, *40*(9), 3308-3322.
57. He, N., Fang, L., & Plaza, A. (2020). Hybrid first and second order attention Unet for building segmentation in remote sensing images. *Science China Information Sciences*, *63*, 1-12.
58. Wang, S., Hou, X., & Zhao, X. (2020). Automatic building extraction from high-resolution aerial imagery via fully convolutional encoder-decoder network with non-local block. *IEEE Access*, *8*, 7313-7322.
59. Liu, Y., Zhou, J., Qi, W., Li, X., Gross, L., Shao, Q., Li, Z. (2020). ARC-Net: An efficient network for building extraction from high-resolution aerial images. *IEEE Access*, *8*, 154997-155010.
60. Zhu, Q., Li, Z., Zhang, Y., & Guan, Q. (2020). Building extraction from high spatial resolution remote sensing images via multiscale-aware and segmentation-prior conditional random fields. *Remote Sensing*, *12*(23), 3983.
61. Zhou, D., Wang, G., He, G., Long, T., Yin, R., Zhang, Z., Luo, B. (2020). Robust building extraction for high spatial resolution remote sensing images with self-attention network. *Sensors*, *20*(24), 7241.
62. Jing, W., Lin, J., & Wang, H. (2020). Building NAS: Automatic designation of efficient neural architectures for building extraction in high-resolution aerial images. *Image Vision Computing*, *103*, 104025.
63. Yang, G., Zhang, Q., & Zhang, G. (2020). EANet: Edge-aware network for the extraction of buildings from aerial images. *Remote Sensing*, *12*(13), 2161.
64. Deng, W., Shi, Q., & Li, J. (2021). Attention-gate-based encoder–decoder network for automatical building extraction. *IEEE Journal of Selected Topics in Applied Earth Observations Remote Sensing*, *14*, 2611-2620.
65. Chen, M., Wu, J., Liu, L., Zhao, W., Tian, F., Shen, Q., Du, R. (2021c). DR-Net: An improved network for building extraction from high resolution remote sensing image. *Remote Sensing*, *13*(2), 294.
66. Chen, Z., Li, D., Fan, W., Guan, H., Wang, C., & Li, J. (2021d). Self-attention in reconstruction bias U-Net for semantic segmentation of building rooftops in optical remote sensing images. *Remote Sensing*, *13*(13), 2524.
67. Chen, K., Zou, Z., & Shi, Z. (2021b). Building extraction from remote sensing images with sparse token transformers. *Remote Sensing*, *13*(21), 4441.
68. Jung, H., Choi, H.-S., & Kang, M. (2022). Boundary enhancement semantic segmentation for building extraction from remote sensed image. *IEEE Transactions on geoscience remote sensing*, *60*, 1-12.
69. Li, Z., Xin, Q., Sun, Y., & Cao, M. (2021). A deep learning-based framework for automated extraction of building footprint polygons from very high-resolution aerial imagery. *Remote Sensing*, *13*(18), 3630.
70. Liu, S., Ye, H., Jin, K., & Cheng, H. (2021). CT-UNet: Context-transfer-UNet for building segmentation in remote sensing images. *Neural Processing Letters*, *53*, 4257-4277.
71. Seong, S., & Choi, J. (2021). Semantic segmentation of urban buildings using a high-resolution network (HRNet) with channel and spatial attention gates. *Remote Sensing*, *13*(16), 3087.
72. Feng, D., Chen, H., Xie, Y., Liu, Z., Liao, Z., Zhu, J., & Zhang, H. (2022). GCCINet: Global feature capture and cross-layer information interaction network for building extraction from remote sensing imagery. *International Journal of Applied Earth Observation Geoinformation*, *114*, 103046.
73. Li, W., Sun, K., Zhao, H., Li, W., Wei, J., & Gao, S. (2022b). Extracting buildings from high-resolution remote sensing images by deep ConvNets equipped with structural-cue-guided feature alignment. *International Journal of Applied Earth Observation Geoinformation*, *113*, 102970.
74. Huang, L., Zhu, J., Qiu, M., Li, X., & Zhu, S. (2022). CA-BASNet: A Building Extraction Network in High Spatial Resolution Remote Sensing Images. *Sustainability*, *14*(18), 11633.
75. Qiu, Y., Wu, F., Yin, J., Liu, C., Gong, X., & Wang, A. (2022). MSL-Net: An efficient network for building extraction from aerial imagery. *Remote Sensing*, *14*(16), 3914.
76. Zhao, H., Zhang, H., & Zheng, X. (2022). A multiscale attention-guided UNet++ with edge constraint for

building extraction from high spatial resolution imagery. *Applied Sciences*, *12*(12), 5960.
77. Xiao, X., Guo, W., Chen, R., Hui, Y., Wang, J., & Zhao, H. (2022). A swin transformer-based encoding booster integrated in u-shaped network for building extraction. *Remote Sensing*, *14*(11), 2611.
78. Yang, S., He, Q., Lim, J. H., & Jeon, G. (2024). RETRACTED ARTICLE: Boundary-guided DCNN for building extraction from high-resolution remote sensing images. *The International Journal of Advanced Manufacturing Technology*, *132*(9), 5171-5171.
79. Yu, M., Zhang, W., Chen, X., Liu, Y., & Niu, J. (2022). An End-to-End Atrous Spatial Pyramid Pooling and Skip-Connections Generative Adversarial Segmentation Network for Building Extraction from High-Resolution Aerial Images. *Applied Sciences*, *12*(10), 5151.
80. Chen, J., Zhang, D., Wu, Y., Chen, Y., & Yan, X. (2022). A context feature enhancement network for building.