

**VISUAL STATE-SPACE BASED MULTI-TASK LEARNING
FOR BUILDING SEGMENTATION AND HEIGHT
ESTIMATION**

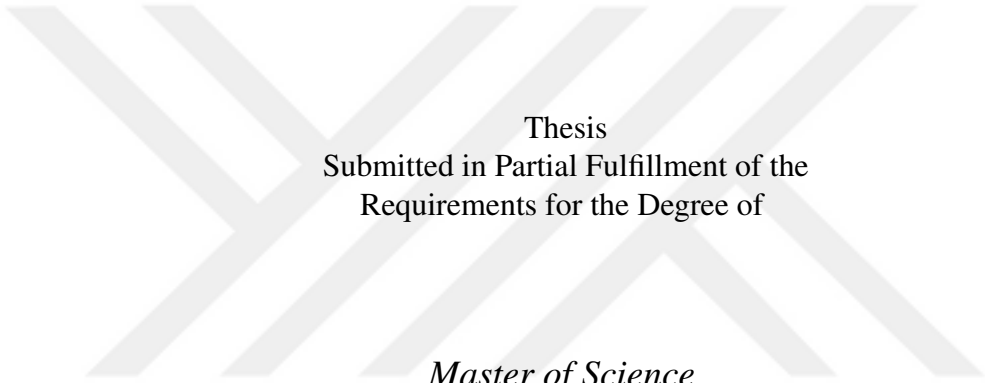


SINAN UTKU ULU

**ÖZYEĞİN UNIVERSITY
DECEMBER, 2025**

VISUAL STATE-SPACE BASED MULTI-TASK LEARNING FOR BUILDING SEGMENTATION AND HEIGHT ESTIMATION

by
Sinan Utku Ulu



Thesis
Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

in
Electrical and Electronics Engineering

Advisor: Prof. Hasan Fehmi Ateş

Graduate School of Science and Engineering
Özyeğin University
İstanbul

December, 2025

VISUAL STATE-SPACE BASED MULTI-TASK LEARNING FOR BUILDING SEGMENTATION AND HEIGHT ESTIMATION

Approved by:

Prof. Hasan Fehmi Ateş, Advisor
Department of Artificial Intelligence and Data
Engineering
Özyeğin University

Asst. Prof. İlknur Karadeniz
Department of Artificial Intelligence and Data
Engineering
Özyeğin University

Prof. Bahadır K. Güntürk
Department of Artificial Intelligence
Engineering
İstanbul Medipol University

Approval Date: November 14, 2025

DECLARATION OF ORIGINALITY

I hereby declare that I am the sole author of this thesis and that this is the true copy of my thesis, including the final revisions, approved by my thesis committee. All data and information have been obtained, produced, and presented in accordance with the rules of research ethics and principles of academic honesty. As required by these rules, to the best of my knowledge I have acknowledged ideas, thoughts, and any copyrighted material in accordance with the standard referencing rules. I certify that any part of this thesis has not been submitted for a degree or diploma in another educational institution.

Sinan Utku Ulu

ABSTRACT

Accurate building segmentation and height estimation from monocular satellite imagery are critical for urban planning, telecommunications, and disaster management. However, conventional deep learning models often struggle with generalization, class imbalance, and underrepresentation of tall buildings. This thesis introduces BuildMamba, a novel multi-task learning framework designed to jointly perform building segmentation and height estimation using only RGB satellite imagery. The model incorporates a visual state-space backbone VMamba for efficient long-range dependency modeling, complemented by a convolutional path for capturing fine-grained local features. To enhance spatial representation, the proposed Mamba Attention Module (MAM) is introduced. For improved height estimation, a boundary-aware Sobel-based weighting strategy is employed to better predict height values at building edges. Extensive experiments on DFC19, DFC23, and Huawei BHE datasets demonstrate the superior performance and generalization ability of BuildMamba compared to competitive baselines. The results show substantial improvements in both segmentation accuracy and height estimation, particularly in complex high-rise urban environments.

Keywords: Remote sensing, multi-task learning, semantic segmentation, building height estimation

ÖZET

Monoküler uydu görüntülerinden doğru bina segmentasyonu ve yükseklik tahmini, kentsel planlama, telekomünikasyon ve afet yönetimi gibi alanlar için büyük önem taşımaktadır. Ancak, geleneksel derin öğrenme modelleri genelleme, sınıf dengesizliği ve yüksek binaların yetersiz temsili gibi sorunlarla sıklıkla karşılaşmaktadır. Bu tez, yalnızca RGB uydu görüntülerini kullanarak bina segmentasyonu ve yükseklik tahminini birlikte gerçekleştirebilen yeni bir çoklu görev öğrenme çerçevesi olan BuildMamba'yı tanıtmaktadır. Model, uzun menzilli bağımlılıkları verimli bir şekilde modelleyebilen görsel durum-uzay tabanlı bir omurga olan VMamba'yı içermekte, buna ek olarak yerel ayrıntıları yakalamak için evrimsel bir yol sunmaktadır. Uzamsal temsili artırmak amacıyla Mamba Attention Module (MAM) önerilmiş ve geliştirilmiştir. Yükseklik tahminini iyileştirmek için ise, bina kenarlarında daha doğru yükseklik tahmini yapılmasını sağlayan, sınır farkındalığına sahip Sobel tabanlı bir ağırlıklandırma stratejisi uygulanmaktadır. DFC19, DFC23 ve Huawei BHE veri kümeleri üzerinde yapılan kapsamlı deneyler, BuildMamba'nın rekabetçi modellere kıyasla üstün performans ve yüksek genelleme yeteneği sergilediğini ortaya koymaktadır. Sonuçlar, özellikle yüksek katlı kentsel alanlar gibi zorlu sahnelerde, hem segmentasyon doğruluğu hem de yükseklik tahmininde önemli iyileşmeler sağlandığını göstermektedir.

Anahtar Kelimeler: Uzaktan algılama, çoklu görev öğrenme, anlamsal bölütleme, bina yüksekliği tahmini

ACKNOWLEDGEMENTS

I am thankful to my advisor, Professor Hasan F. Ateş for his invaluable support and patience with me. His guidance and mentorship helped me a lot in my academic development. Special thanks to my father, who worked day and night to provide me with a good future. To my brother, for his constant understanding, thoughtfulness, and unwavering support. To my mother, for her endless sacrifices. Finally, special thanks to Ece, who always manages to motivate me with her great energy. I would also like to thank Huawei Türkiye R&D Center for their support and for providing the Huawei BHE dataset used in this study.

This thesis is supported in part by TÜBİTAK Grant - No: 5230108.

TABLE OF CONTENTS

DECLARATION OF ORIGINALITY	iii
ABSTRACT	iv
ÖZET	v
ACKNOWLEDGEMENTS	vi
LIST OF TABLES	ix
LIST OF FIGURES	x
LIST OF ACRONYMS AND ABBREVIATIONS	xiii
1 INTRODUCTION	1
1.1 Motivation	2
1.2 Concepts	5
1.2.1 Remote Sensing	5
1.2.2 Semantic Segmentation	6
1.2.3 Height Estimation	6
1.2.4 Multi-Task Learning	7
1.3 Contributions	7
1.4 Overview of the Thesis	8
2 LITERATURE REVIEW	10
2.1 Monocular Height Estimation	10
2.2 Semantic Segmentation	13
2.3 Multi-Task Learning	17
3 METHODOLOGY	21
3.1 Overview of BuildMamba Architecture	21
3.2 VMamba	24
3.2.1 Architecture Overview	25
3.2.2 Selective State-Space Models	25

3.3	Dual-Path Encoder	27
3.4	Mamba Attention Module (MAM)	28
3.5	FPN	29
3.6	Semantic Segmentation Head	30
3.7	Height Estimation Head	32
3.8	Multi-Task Learning	34
3.9	Architectural and Loss Function Variants	34
3.9.1	Backbone	35
3.9.2	Heads	37
3.9.3	Loss Functions	42
4	EXPERIMENTS AND RESULTS	45
4.1	Datasets	45
4.1.1	DFC19	45
4.1.2	DFC23	47
4.1.3	Huawei BHE	47
4.1.4	Dataset Comparison	48
4.2	Implementation Details	50
4.3	Metrics	50
4.4	Results	53
4.5	Discussion and Ablation Study	65
4.5.1	Impact of Proposed Modules on Model Performance	66
4.5.2	Loss Function Variants	69
4.5.3	Comparison of Backbones	69
4.5.4	Comparison of Different Segmentation Heads	71
4.5.5	Comparison of Height Estimation Heads	72
5	CONCLUSION AND FUTURE WORK	74
	REFERENCES	76

LIST OF TABLES

Table 4.1: Dataset statistics for the DFC19 dataset. The table summarizes the number of images and buildings, as well as key numerical statistics relevant to building height analysis and model development.	47
Table 4.2: Dataset statistics for the DFC23 dataset. The table summarizes the number of images and buildings, as well as key numerical statistics relevant to building height analysis and model development.	48
Table 4.3: Dataset statistics for the Huawei BHE dataset. The table summarizes the number of images and buildings, as well as key numerical statistics relevant to building height analysis and model development.	48
Table 4.4: Semantic Segmentation Performance Across Models on DFC19 Dataset	54
Table 4.5: Semantic Segmentation Performance Across Models on DFC23 Dataset	55
Table 4.6: Semantic Segmentation Performance Across Models on Huawei BHE Dataset.....	56
Table 4.7: Building Height Estimation Performance on DFC19 Dataset.	61
Table 4.8: Building Height Estimation Performance on DFC23 Dataset.	61
Table 4.9: Building Height Estimation Performance on Huawei BHE Dataset.	62
Table 4.10: Configuration overview of models based on MAM, Adaptive Binning, and Boundary-Aware Loss.....	67
Table 4.11: Ablation study on the impact of MAM, Adaptive Binning, and Loss Functions on DFC23 dataset.....	68
Table 4.12: Ablation study on the impact of Different Height Loss Functions on DFC23 dataset.....	69
Table 4.13: Comparison of different Backbone architectures on the DFC23 dataset. .	70
Table 4.14: Comparison of Different Segmentation Heads on the DFC23 Dataset. ...	72
Table 4.15: Comparison of Different Height Heads on the DFC23 dataset.	73

LIST OF FIGURES

Figure 1.1:	The challenge of reconstructing 3D information, specifically building segmentation and pixel-wise height estimation from single-view satellite imagery. Given only an RGB image captured from a satellite, the goal is to generate both a segmentation mask to identify built structures and a corresponding height map to estimate elevation. These outputs contribute to DSM generation, enabling 3D reconstruction in the absence of LiDAR or multi-view data. Accurate DSMs are critical for various real-world applications, including urban planning, disaster management, agriculture, and telecommunications.	5
Figure 3.1:	The proposed BuildMamba architecture. It consists of a dual-path encoder that fuses global features from the VMamba backbone and local features from CNN layers. The architecture includes a Mamba Attention Module (MAM), and a Feature Pyramid Network (FPN), followed by task-specific decoders for semantic segmentation and height estimation.	22
Figure 3.2:	Overall architecture of VMamba [1]. The model begins with a patch embedding layer that tokenizes the input image into non-overlapping patches. These tokens are passed through multiple stages of Mamba blocks, each consisting of a state-space-based recurrent update mechanism designed to capture long-range spatial dependencies efficiently. Each stage also includes layer normalization and MLP blocks to stabilize and enhance feature learning. The architecture produces multi-scale feature maps.	26
Figure 3.3:	Architecture of the Mamba Attention Module (MAM). The input feature map $F_{\text{in}} \in \mathbb{R}^{C \times H \times W}$ is processed via two parallel spatial pathways: one applies adaptive average pooling along the height dimension ($H \times 1$), and the other along the width dimension ($1 \times W$). Each pooled output is passed through a fully connected layer to produce intermediate representations. These are combined via element-wise multiplication, followed by a softmax normalization to generate an attention map. This map is used to reweight the input features, which are then added back to the original feature map through a residual connection scaled by a learnable parameter α , resulting in the refined output F_{out}	28

- Figure 3.4:** Illustration of the Feature Pyramid Network (FPN) [2] architecture. The backbone produces hierarchical feature maps from Stage 1 to Stage 4, with decreasing spatial resolution and increasing semantic richness. These features are connected to the feature pyramid via lateral 1×1 convolutions. A top-down pathway upsamples higher-level features and merges them with lower-level lateral features via element-wise addition, producing a set of multi-scale representations $\{P_1, P_2, P_3, P_4\}$. This structure enables the model to utilize both fine and coarse spatial features for downstream tasks. 30
- Figure 3.5:** Overview of the BuildFormer [3] Block architecture. The first path begins with Batch Normalization followed by a Window-based Linear Multi-Head Self-Attention (W-LMHSA) module. This output is added back to the input via a residual connection. The second stage consists of another Batch Normalization layer followed by a Convolutional Multi-Layer Perceptron (C-MLP), responsible for enhancing channel-wise feature interactions. A final residual connection ensures gradient flow and stabilizes training. 36
- Figure 3.6:** Overview of the Swin Transformer architecture [4]. The model processes the input image through a hierarchical structure consisting of four stages. Each stage includes multiple Swin Transformer blocks that operate within shifted local windows to perform self-attention efficiently. Between stages, patch merging operations downsample spatial dimensions while increasing feature depth, enabling multi-scale feature extraction. This hierarchical design allows Swin Transformer to balance local detail preservation with global context modeling 37
- Figure 3.7:** Overview of the AdaBins module [5] for hybrid regression in depth, or height estimation in this case. AdaBins module combines classification and regression in a hybrid formulation to produce continuous height outputs. The feature representation R generated by the mViT backbone is passed through a convolutional layer to predict class probabilities over N discretized height bins. A softmax is applied to obtain per-pixel class distributions. Simultaneously, adaptive bin centers are computed based on learnable bin widths b and a weighted sum is performed to regress the final height value at each pixel. This approach allows the network to discretize the continuous height space into bins. 39

Figure 3.8:	Overview of the UPerNet architecture [6] for dense prediction tasks, adapted here for semantic segmentation. UPerNet leverages a unified framework built upon a Feature Pyramid Network to integrate multi-scale features extracted from the backbone network. Specifically, features from different stages (e.g., $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, $\frac{1}{32}$) are first aligned through lateral connections and then fused to obtain a semantically enriched representation. PPM is applied to the deepest feature to capture global context, which is then merged with the upsampled features. Finally, the aggregated feature map is fed into a lightweight decoder head for semantic segmentation. This architecture enables effective fusion of both local spatial details and global contextual information across scales.	41
Figure 3.9:	Overview of the Atrous Spatial Pyramid Pooling (ASPP) module [7], which is used as an alternative segmentation head to analyze the performance of the proposed model. ASPP module enhances the receptive field by applying multiple parallel atrous convolutions with different dilation rates to the same input feature map. Each branch captures contextual information at a distinct scale, allowing the network to effectively integrate both local details and long-range dependencies. The outputs from all branches are concatenated and fused through a 1×1 convolution followed by batch normalization and activation to produce the final feature representation. This structure enables multi-scale context aggregation and ability to segment complex urban structures in satellite imagery.	43
Figure 4.1:	Visual Comparison of Images in Each Dataset.	46
Figure 4.2:	Visualization of predicted building segmentations on the DFC19 dataset, comparing results across competing models.	57
Figure 4.3:	Visualization of predicted building segmentations on the Huawei BHE dataset, comparing results across competing models.	58
Figure 4.4:	Visualization of predicted building segmentations on the DFC23 dataset, comparing results across competing models.	59
Figure 4.5:	Visualization of predicted building heights on the DFC19 dataset, comparing results across competing models.	63
Figure 4.6:	Visualization of predicted building heights on the DFC23 dataset, comparing results across competing models.	64
Figure 4.7:	Visualization of predicted building heights on the Huawei BHE dataset, comparing results across competing models.	65

LIST OF ACRONYMS AND ABBREVIATIONS

ASPP	Atrous Spatial Pyramid Pooling
BHE	Building Height Estimation
CNNs	Convolutional Neural Networks
DSM	Digital Surface Model
FPN	Feature Pyramid Network
IoU	Intersection over Union
LiDAR	Light Detection and Ranging
MTL	Multi-Task Learning
MLP	Multi-Layer Perceptron
nDSMs	Normalized Digital Surface Models
PPM	Pyramid Pooling Module
RGB	Red, Green, Blue
RMSE	Root Mean Square Error
SAR	Synthetic Aperture Radar
ViTs	Vision Transformers

1. INTRODUCTION

Remote sensing has emerged as a fundamental technology for observing and analyzing the Earth’s surface by capturing data from satellite or airborne sensors. It enables the large-scale, timely, and cost-effective collection of geospatial information without direct contact, making it indispensable for a variety of applications such as land cover classification, urban planning, environmental monitoring, disaster management, and agricultural assessment. Moreover, Earth vision includes various vision tasks such as change detection, object detection, segmentation, and regression, which support applications such as urban planning, agriculture, land cover mapping, and disaster management. In this context, semantic segmentation and height estimation are essential tasks, as these applications require 3D information about the environment. Building segmentation and height estimation provide both semantic and geometric understanding of urban environments. Semantic segmentation of buildings enables the tracing of their spatial extent and shape, while height estimation provides significant 3D structural information, contributing to the reconstruction of 3D cities. These capabilities are particularly useful in high-density urban planning, infrastructure development, and hazard risk assessment.

Among existing height extraction methods, Light Detection and Ranging (LiDAR) [8, 9] is highly accurate but costly and resource-intensive. Single Red, Green, Blue (RGB) satellite imagery provides a practical and scalable alternative, being widely available and suitable for large-scale applications. However, estimating height from a single monocular RGB satellite image remains a significant challenge due to the lack of direct height information and factors like viewing angle, resolution, and atmospheric conditions. Therefore, advanced algorithms are required for remote sensing imagery applications to make it more scalable and cheap for real-world applications. For instance, Copernicus-FM [10], a model for downstream applications in remote sensing, can utilize both spectral and non-spectral inputs, and incorporates metadata with Fourier encoding to form a foundation

model for Earth observation. However, to get the most out of it, spectral and non-spectral inputs should be used together, which is not always possible in real life.

1.1 Motivation

Accurate 3D reconstruction of terrain and surface features, especially through the generation of DSM (Digital Surface Model) has become a foundational task in remote sensing, computer vision, and geospatial analysis. DSM creation captures the full geometric structure of a region, including buildings, vegetation, and topographic variations, enabling a comprehensive representation of real-world surfaces. This level of spatial detail has applications across multiple domains. In urban planning, DSMs facilitate shadow analysis, line-of-sight studies, and smart city development. In environmental monitoring, they enable flood risk assessment, forest canopy modeling, and erosion tracking. Moreover, industries like telecommunications benefit from DSMs for network planning and signal propagation modeling, while autonomous navigation systems rely on 3D environmental understanding for obstacle avoidance and path planning. Approaching the problem from a DSM-generation perspective thus not only enhances the fidelity of spatial data but also unlocks a wide range of high-impact, real-world applications.

Furthermore, accurate building height estimation plays a crucial role in the design of 5G networks, which rely heavily on line-of-sight and elevation-aware signal propagation models. As 5G antennas and base stations are sensitive to occlusions and multipath effects caused by urban structures, having precise knowledge of building heights allows for more efficient placement of network infrastructure, improved signal coverage, and minimized interference. In this context, extracting reliable semantic and height information from only optical satellite imagery without depending on expensive LiDAR or Synthetic Aperture Radar (SAR) data has become increasingly important for scalable deployment. Advanced deep learning techniques, particularly in Multi-Task Learning (MTL)

frameworks, enable the joint semantic segmentation and height estimation of buildings from RGB images, making such models highly practical for real-world applications and the future of urban connectivity.

Digital Surface Model (DSM) generation from satellite imagery has long been a central problem in remote sensing, urban reconstruction, and 3D geographic information systems. Traditional approaches rely on LiDAR or stereo imagery to extract height information, offering high accuracy but at the cost of expensive equipment, complex setups, and limited applicability in certain environments. With the growing availability of monocular RGB imagery from satellites, UAVs, and surveillance systems, there is increasing demand for cost-effective and flexible methods for height estimation and DSM generation. These 3D reconstructions have critical applications in urban planning, infrastructure monitoring, environmental modeling, and disaster response. Recent advances in deep learning have enabled direct pixel-wise height estimation from single monocular images, offering a lightweight alternative to traditional methods. In this thesis, a multi-task learning network is proposed to jointly perform building segmentation and height estimation from single-view satellite imagery in complex urban environments (see Figure 1.1).

Pixel-wise height estimation from monocular images is a crucial task in remote sensing and urban reconstruction. Unlike LiDAR or stereo imaging methods, monocular approaches offer lower cost and greater deployment flexibility, especially in scenarios where only RGB imagery is available, such as satellite, UAV, or surveillance applications. Height estimation is essential for creating 3D models, infrastructure monitoring, and disaster response planning. Recent deep learning methods have made it possible to infer height from single RGB images. However, this task remains fundamentally ill-posed, as height cues are limited and highly dependent on scene context. In this thesis, a multi-task learning network is proposed such that it benefits from semantic priors to improve

height prediction accuracy, especially for challenging urban environments. Furthermore, pixel-wise height estimation presents challenges when applied to monocular imagery, particularly in complex urban scenes and images with low resolution. One major issue is assigning a continuous height value to every pixel in the image, regardless of whether it belongs to a building or the background. In practice, this results in a significant class imbalance: background pixels vastly outnumber building pixels, resulting in the model being prone to learning a distribution biased toward background characteristics, which leads to the underestimation of tall buildings. This is worsened by the fact that tall buildings often occupy a smaller footprint and are surrounded by background regions, making it difficult for the network to isolate and preserve their height characteristics. To mitigate this, recent approaches like HTC-DC Net [11] and HENet [12] introduce a classification-regression paradigm. Inspired by the AdaBins [5] module, these models discretize height ranges into bins (essentially treating height as a classification problem) and then refine the prediction via regression within each bin. However, while this method indicates improvements in overall performance, its effectiveness on tall buildings or skyscrapers remains speculative.

Another critical issue in pixel-wise height estimation arises at building boundaries. The prediction quality often deteriorates at the edges of buildings. This boundary degradation can be attributed to several factors, including interpolation artifacts, receptive field limitations, and misalignment in semantic context. Essentially, as we move away from the building’s geometric center, the contextual clues available to the model become noisier or weaker, leading to spatial inconsistency in height prediction—even though all pixels should ideally be assigned the same height. To address this, some works, such as [13], have proposed edge-aware loss weighting schemes. By applying a Sobel operator to the segmentation mask or ground truth height map, the model explicitly identifies boundary regions and assigns higher loss weights to these pixels during training. This encourages

the network to pay greater attention to edges, improving prediction consistency across the entire object and ensuring sharper, more accurate boundaries in the output height maps. The technique promotes balanced learning between the interior and edge regions of buildings, effectively reducing artifacts and promoting height uniformity.

This thesis addresses these and several other challenges in pixel-level height estimation and offers an efficient deep learning solution that achieves state-of-the-art performance in both building segmentation and height estimation across several datasets.

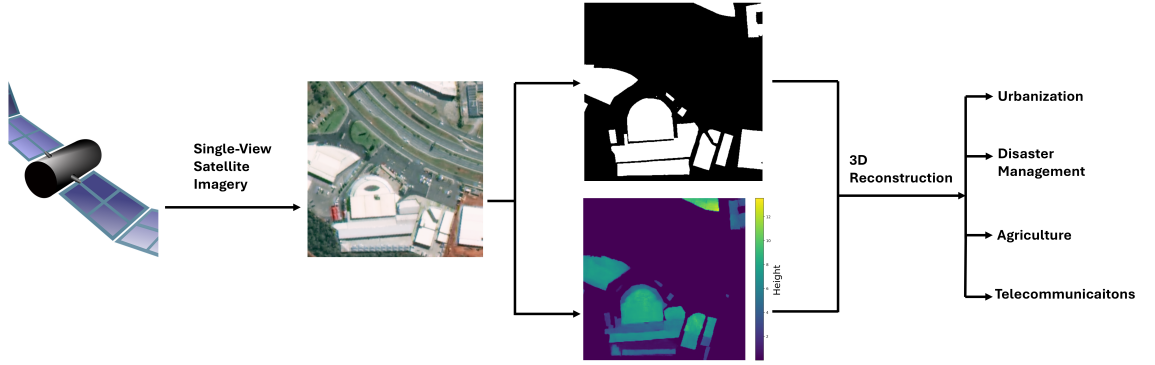


Figure 1.1: *The challenge of reconstructing 3D information, specifically building segmentation and pixel-wise height estimation from single-view satellite imagery. Given only an RGB image captured from a satellite, the goal is to generate both a segmentation mask to identify built structures and a corresponding height map to estimate elevation. These outputs contribute to DSM generation, enabling 3D reconstruction in the absence of LiDAR or multi-view data. Accurate DSMs are critical for various real-world applications, including urban planning, disaster management, agriculture, and telecommunications.*

1.2 Concepts

1.2.1 Remote Sensing

Remote sensing [14] refers to the process of acquiring information about objects or areas from a distance, typically using satellite or aerial platforms. It enables the observation of Earth’s surface without direct contact, allowing for wide-area and frequent data collection. Data from remote sensing can be gathered utilizing different types of sensors,

such as optical, thermal, multispectral, hyperspectral, and SAR. In urban contexts, remote sensing is instrumental for tasks such as land-use classification, infrastructure mapping, environmental monitoring, and disaster response. The data's scalability and repeatability make it an indispensable tool in modern geospatial analysis and smart city development.

1.2.2 Semantic Segmentation

Semantic segmentation [15] is a task in computer vision that requires making dense predictions by assigning a category label to every pixel within an image. In remote sensing, semantic segmentation is widely used to extract meaningful regions such as buildings, roads, vegetation, or water bodies from high-resolution aerial or satellite images. The task is particularly valuable for applications like land cover mapping, urban planning, and cadastral management. Accurate building segmentation, in particular, enables precise footprint detection which is crucial for 3D cities. The challenge lies in handling varying object sizes, occlusions, and class imbalance, all of which are common in overhead imagery.

1.2.3 Height Estimation

Building height estimation from remote sensing imagery is a critical task for creating 3D representations of urban landscapes. It typically involves predicting the vertical structure of buildings using input modalities such as stereo imagery, LiDAR, SAR, or even single RGB images. Height information supports various applications including urban simulation, solar potential estimation, and wireless communication planning such as 5G deployment. In the context of monocular imagery, height estimation becomes a challenging regression problem due to the lack of direct depth cues. Deep learning approaches address this by extracting semantic and structural patterns that correlate with vertical geometry, sometimes using auxiliary information such as building shadows [16] or footprints.

1.2.4 Multi-Task Learning

In multi-task learning (MTL) [17] one model is trained to learn several tasks at the same time. The key idea is to leverage shared representations between tasks to improve generalization and learning efficiency. In remote sensing, combining tasks such as semantic segmentation and height estimation under a unified framework is particularly effective, because both rely on spatial understanding and context. While segmentation delineates object boundaries, height estimation refines the model’s sense of object structure and scale. By training these tasks together, MTL promotes inductive transfer between tasks, often resulting in improved accuracy and robustness compared to training separate models. Furthermore, it reduces computational cost and inference time, making it well-suited for real-world large-scale remote sensing applications.

1.3 Contributions

This thesis is about building segmentation and height estimation from a single RGB satellite image. The motivation for this work lies in tackling key challenges observed in prior literature, such as the lack of multi-modal data, inaccurate boundary predictions, and difficulties in generalizing to different high-resolution urban data. Additionally, this study addresses the underestimation of tall buildings and background pixel domination, which are common in monocular-based remote sensing models.

To overcome these limitations, the proposed method leverages recent advances in deep learning architectures—specifically visual state-space modeling via the VMamba [1] encoder and introduces the Mamba Attention Module (MAM), which enhances spatial relationship modeling by injecting localized context into global descriptors. This work also seeks to remove dependency on costly auxiliary data sources such as LiDAR or SAR, thereby promoting the feasibility of high-quality mapping using only RGB imagery.

Moreover, a novel boundary-aware L1 loss is incorporated to model training,

which aligns with the motivation to achieve fine-grained precision around edges, which is critical for realistic urban reconstruction and telecommunications applications such as 5G signal planning. Thus, the motivation behind this research is not only to improve performance metrics but also to redefine the feasibility of high-quality 3D urban mapping using accessible RGB imagery alone.

The main contributions of this research are as follows:

- BuildMamba introduces a boundary-aware L1 loss to improve height estimation near building boundaries.
- The model employs the VMamba encoder for efficient global-local modeling with linear computational complexity, addressing scalability challenges.
- BuildMamba integrates the MAM within a dual-path backbone, enhancing feature representation and multi-task learning performance.
- New Huawei BHE dataset offers various types of cities with different data characteristics on each city and large dataset size. BuildMamba achieves state-of-the-art results on both building segmentation and height estimation across this and other benchmarks.
- Various extensions of the model and its multi-task framework are developed and tested in the context of building segmentation and height estimation.

1.4 Overview of the Thesis

This thesis presents a novel multi-task learning framework, BuildMamba, designed for simultaneous building segmentation and height estimation from monocular satellite imagery. The proposed model leverages a VMamba [1] backbone and task-specific decoders to effectively learn both semantic and geometric cues from urban scenes.

The work is motivated by the increasing need for accurate and scalable urban city modeling, supporting applications such as 5G network deployment, tall building height estimation, and other remote sensing tasks where reliance on expensive modalities like SAR or LiDAR is impractical.

The rest of the thesis is structured as follows: Chapter 2 presents a comprehensive review of literature, covering recent developments in monocular height estimation, semantic segmentation, and multi-task learning approaches. Chapter 3 details the proposed methodology, including the architectural components, loss formulations, and training strategy. Chapter 4 discusses the experimental setup, datasets, and evaluation metrics, followed by an in-depth analysis of results, including ablation studies. Finally, Chapter 5 concludes the thesis with key findings, limitations, and directions for future research.

2. LITERATURE REVIEW

In this thesis, the literature review is structured into three main categories: monocular height estimation (including depth estimation, due to their methodological similarity in deep learning), semantic segmentation, and multi-task learning.

2.1 Monocular Height Estimation

Estimating height from monocular remote sensing imagery presents significant challenges; several methods are suggested to enhance the data and boost the learning abilities of the models. Specific strategies integrate extra modalities, such as SAR images, which offer additional structural insights under various imaging conditions. Other methods [18, 16] exploit shadow cues to infer height more accurately. [18] propose a physically interpretable approach for building height estimation from remote sensing images by leveraging shadow and side information. The method begins by detecting individual buildings, after which corresponding shadow and side regions are extracted using geometric rules and solar azimuth data. From these regions, hand-crafted features such as geometric ratios, intensity, and texture are computed. These features are then used to train a regression model to predict building heights. These auxiliary informations help mitigate the limitations of monocular RGB imagery and improve the robustness of height prediction models. IM2HEIGHT [19] uses a convolutional-deconvolutional network to generate height maps, while IM2ELEVATION [20] combines LiDAR and optical data to address misregistration artifacts.

DORN [21] introduces a new approach to monocular depth estimation, a similar task to height estimation, by formulating it as an ordinal regression problem rather than predicting a continuous value. The method discretizes the continuous depth values into intervals and predicts the depth class at each pixel utilizing an ordinal regression network. It allows the model to capture relative ordering between depths more effectively.

3DCentripetalNet [22] estimates building height from monocular satellite images using a center-based detection framework extended into 3D. The model treats each building as a 3D object and predicts its center point, footprint, and height simultaneously. One of the components is the 3D centripetal loss, which enhances the geometric coherence between the predicted 2D footprint and the height by pulling predicted center points toward the geometric center of the building in 3D space. This approach allows the network to model building geometry more holistically compared to pixel-wise regression.

Adabins [5] proposes a method that formulates the task as classification followed by regression, utilizing adaptive depth binning. Instead of using fixed bin strategies, AdaBins predicts the optimal bin edges dynamically for each input image. The network predicts depth by classifying each pixel into one of the adaptive bins, then applies regression to continuous depth values through a weighted combination of bin centers. This adaptive binning mechanism allows the model to allocate finer resolution to important depth ranges.

BinsFormer [23] is a transformer-based refinement of the adaptive binning strategy [5] for monocular depth estimation. BinsFormer introduces a query-based transformer architecture where each learnable query corresponds to an adaptive depth bin. The model separates the task of predicting bin centers from pixel-wise bin classification, enabling a transformer decoder to provide refined bin centers and align them with spatial locations. Moreover, cross-attention between feature maps and bin queries captures better spatial details, resulting in sharper depth boundaries and enhanced accuracy in complex scenes. The method demonstrates the benefits of combining adaptive discretization with transformer-based representation learning for monocular depth prediction.

HENet [12] integrates [5] to design a classification-regression paradigm for height estimation. The proposed method utilizes a roughly generated height map, which is generated by the decoder and refined by segmentation-based AdaBins, to achieve improved

height regression results. Moreover, [24] is used in [12] to filter out background introduced in segmentation based AdaBins.

According to [11], one major challenge for height estimation is the long-tailed distribution of height values where tall building pixels are sparse, while background pixels dominate, often leading regression-based models to underperform on high-rise structures. To overcome this, HTC-DC Net [11] proposes a hybrid classification-regression framework which is inspired by [5]. Their method learns adaptive bin edges and bin probabilities using a Head-Tail Cut mechanism, which separates foreground and background pixels to better handle the long-tailed distribution. The final height predictions are obtained by computing a weighted average of bin centers, and probabilistic modeling is employed to regularize bin probabilities, enhancing prediction stability and accuracy.

LUMNet [13] designs a network that leverages from land use knowledge to guide a multiscale height estimation network. Instead of using MTL network, the model is guided by semantic labels to incorporate height estimations.

HeightFormer [25] introduces a transformer-based framework for single-imagery height estimation that effectively balances global and local feature understanding. Built upon a multiscale vision transformer backbone, it integrates a novel bilateral feature pyramid fusion mechanism comprising a Stepwise Fusion Module and a Multiscale Fusion Module to progressively refine feature representations. These modules allow HeightFormer to preserve fine-grained geometric details while maintaining global contextual awareness, a common limitation in CNN-based single-imagery height estimation methods. Furthermore, the proposed Heightbins module converts height regression into a classification problem using adaptive bin widths and attention maps, enabling robust and precise estimation of building heights.

AMNet [26] introduces a transformer-inspired encoder-decoder framework for

single-image building height estimation, designed to enhance multi-scale feature representation and detail recovery. The encoder integrates a Flexible Attention Mechanism that combines deformable convolutions with multi-head self-attention to capture both local geometric variations and long-range dependencies. To better balance global and local contexts, a Dense Atrous Spatial Pyramid Pooling module aggregates multi-scale features through dense connections, improving feature flow and semantic consistency. In the decoder, the Progressive Fusion Module aligns and fuses high-resolution local features with low-resolution global features via learnable upsampling and residual refinement, enabling precise reconstruction of fine details.

[27] presents an instance-level multi-task learning framework designed to infer building heights directly from off-nadir satellite images, where the roof-to-footprint displacement naturally encodes geometric height cues. Unlike traditional ortho-rectified or pixel-level methods, this approach leverages instance segmentation and geometric proxies to capture spatial relationships among buildings and reduce ambiguity in height estimation.

[28] proposes a novel semantic-guided unsupervised domain adaptation framework to estimate building heights from single-view ortho-photos without relying on labeled Normalized Digital Surface Models (nDSMs) data in the target domain. The method addresses two critical challenges in height estimation under domain shift: maintaining semantic consistency during unpaired image translation and capturing multi-task correlations between semantic segmentation and height estimation

2.2 Semantic Segmentation

In remote sensing, semantic segmentation plays a crucial role, where each pixel in an image is classified into one of several semantic categories. This pixel-level image understanding is making a impact on urban analysis and land cover classification. Recent

progress in deep learning has enhanced segmentation performance by effectively capturing both fine-grained textures and broader contextual information. These developments also affect height estimation indirectly.

The effect of self-supervised learning methods also influences segmentation height estimation problems. [29] suggests improving semantic segmentation performance in remote sensing by leveraging monocular height estimation as a self-supervised pretext task during training.

SegFormer [30] presents a semantic segmentation framework that addresses the shortcomings of current transformer-based and CNN-based models by proposing a hierarchical transformer encoder with a MLP decoder. The fundamental concept is to use transformers to capture long-range dependencies while ensuring computational efficiency through effective design.

BuildFormer [3] introduces linear complexity attention mechanism to address the high computational cost of traditional self-attention mechanism and aims to improve building extraction performance from remote sensing images by incorporating vision transformers.

FarSeg and FarSeg++ [31, 32] is designed explicitly for domain generalization to enhance semantic segmentation performance in remote sensing imagery, particularly under domain shifts and across diverse geographical distributions, by combining deformable convolutions and Swin Transformer blocks.

[33] proposes a parallel architecture that combines the Swin Transformer and U-Net to improve semantic segmentation performance in remote sensing imagery. Swin Transformer blocks capture global context and long-range dependencies while U-Net captures the local features. At the same time, the U-Net-inspired decoder with skip connections ensures effective recovery of spatial details, unlike traditional pooling, patch merging, and expanding mechanisms used for downsampling and upsampling.

PF-SegNet [34] also combines multiple encoder paths at different scales to capture features at specific resolutions, enabling the network to effectively capture multi-scale contextual information, particularly in preserving object boundaries and structural integrity in complex scenes for semantic segmentation in remote sensing imagery.

UANet [35] introduces a framework that explicitly models and mitigates uncertainty during building footprint extraction from high-resolution remote sensing imagery. Traditional encoder–decoder architectures often suffer from inconsistent predictions due to variations in building size, material, and illumination conditions. To overcome these issues, UANet integrates a Prior Information Guide Module and an Uncertainty-Aware Fusion Module to enhance both spatial detail and confidence in predictions. The Prior Information Guide Module leverages cross-attention guided by an initial uncertainty map to refine semantic features, while the Uncertainty-Aware Fusion Module adaptively fuses multi-scale information based on an Uncertainty Rank Algorithm that categorizes pixels by their uncertainty levels. Together, these mechanisms enable the network to progressively focus on ambiguous regions and suppress uncertainty in a top-down manner.

CBR-Net [36] presents an approach for building height estimation from single-view remote sensing images by emphasizing spatial and contextual relationships among buildings. Instead of treating each pixel independently, CBR-Net captures inter-building dependencies that influence relative height perception in urban environments. The model employs a dual-stream architecture, combining a contextual relation stream for extracting global spatial interactions with a boundary refinement stream that preserves sharp building edges and shape details. These two branches are fused adaptively through an attention-based mechanism, ensuring balanced learning of semantic context and geometric precision.

MSHFormer [37] is a hybrid architecture that fuses the local detail-capturing ability of Convolutional Neural Networks (CNNs) with the global modeling power of transformers. The model incorporates a Multiscale Local Perception module to enhance multi-scale feature representation, an Edge Enhancement Module to refine building boundaries and suppress background noise, and a Group Alignment Feature Fusion Module to align encoder and decoder features for better integration.

BCTNet [38] introduces a bi-branch cross-fusion transformer architecture for precise building footprint extraction from high-resolution satellite images, showcasing notable improvements in semantic segmentation.

[39] proposes an approach called decoupled encoder network to improve building footprint extraction from remote sensing images by explicitly decoupling the learning of semantic and edge representations. Traditional segmentation models often struggle to balance accurate region prediction with precise boundary delineation. DE-Net addresses this by using a shared Resnet based encoder followed by two distinct decoder branches: one for semantic features to capture the interior regions of buildings and another for edge features to detect structural boundaries. This separation allows each branch to specialize in its task. The outputs of these branches are then fused using a feature interaction module, enabling joint reasoning over both semantic and edge cues to produce refined segmentation results.

[40] suggests a novel dual-stream architecture that simultaneously leverages local fine-grained details and global contextual features for accurate building footprint extraction. The local stream captures detailed spatial information using a lightweight CNN-based encoder-decoder structures. In parallel, the global stream adopts a transformer-based encoder that captures long-range dependencies and global semantic context, which helps in understanding complex scenes with diverse building distributions. To effectively

integrate both streams, a cross-fusion module is employed, which enables dynamic information exchange between the local and global branches. Furthermore, a multi-scale feature fusion module is designed to refine predictions by combining hierarchical features at different resolutions. The final segmentation map is obtained through a fusion decoder that integrates the complementary information learned by both streams.

Vision Transformers (ViTs) [41] advance segmentation by capturing global dependencies, with methods like Swin Transformer [4] using shifted windows for scalability. On the other hand, VMamba [1] introduces a novel visual state space model that combines the strengths of state space models and convolutional architectures for efficient and scalable visual representation learning. VMamba utilizes a long convolution operator instead of conventional transformer, which uses attention mechanisms with quadratic complexity, to capture both local and global dependencies with linear time and space complexity. With its selective-scan mechanism, the model selectively processes parts of the input image tokens, adapting the scope of information scanned based on content, rather than applying uniform attention or convolution. RS3Mamba [42] adapts VMamba [1] to remote sensing tasks for semantic segmentation by proposing a dual-branch network and utilizing visual state space as an auxiliary branch.

2.3 Multi-Task Learning

Since semantic segmentation and height regression can mutually influence and constrain each other, adopting a multi-task learning framework offers promising improvements in height estimation performance. PopNet [43] is a multi-task learning network for semantic segmentation and height estimation, utilizing an Encoder–Dual Decoder architecture where the first decoder generates a semantic feature pyramid. At the same time, the second decoder constructs a height estimation pyramid. It predicts heights with an anchor-based approach, which benefits from class-specific statistical height distributions

to predict normalized DSM values more robustly. Instead of directly regressing absolute height, the model estimates scale offsets relative to class-specific anchors.

DSMNet [44] addresses the height estimation problem differently by employing a two-stage network. The first stage is a multi-task learning network with three branches that perform height prediction, semantic segmentation, and an auxiliary task as surface normal guidance. Furthermore, in the second stage, a denoising autoencoder is utilized to refine the predictions from stage 1.

[45] presents a method for reconstructing 3D building structures by forcing boundary and structural awareness for the building height estimation process. Their method employs a boundary-aware decoder that combines both semantic segmentation and edge detection feature maps to improve the geometric precision of building outlines and height predictions. Another contribution is the height-gradient loss, which penalizes height inconsistencies at building edges, encouraging sharper and more structurally coherent height maps.

LIGHT [46] unifies building instance segmentation and pixel-wise height estimation into a single end-to-end multi-task learning network by proposing gated cross-task interaction, enhancing mutual feature learning between instance segmentation and height estimation branches.

MFTSC [47] is a multi-task learning network designed for estimating building heights and semantic segmentation of buildings. The method utilizes multi-level features with semantic guidance to enhance the accuracy and spatial consistency of height prediction. Additionally, a structure-aware loss is used to enforce spatial smoothness and preserve the 3D morphology of buildings. The approach improves the height estimation of high-rise structures and reduces artifacts caused by background clutter.

HGDNet [48] is dual-decoder framework for joint building footprint extraction

and height estimation from remote sensing imagery. The model introduces a height-hierarchical segmentation branch that leverages discrete height categories derived from normalized nDSMs to guide the height estimation process. By converting height labels into semantic classes, such as low-rise, mid-rise, and high-rise buildings, this auxiliary branch serves as a constraint that improves the accuracy of the main height regression branch. In addition, because of the uneven distribution of building heights, height maps are processed using a logarithmic normalization to facilitate faster convergence and a more normal distribution.

[49] proposes a framework for the joint learning of semantic segmentation and building height estimation, enhanced by contrastive learning. To improve representation learning, a contrastive loss is used to align features of similar semantic and height categories while pushing apart dissimilar ones in the feature space. This approach pushes the backbone to generate more diverse and semantically meaningful feature maps, which helps both tasks. Additionally, auxiliary contrastive heads are attached to intermediate layers to apply contrastive supervision at multiple scales.

To summarize, the reviewed literature is categorized into three domains: monocular height estimation, semantic segmentation, and multi-task learning. Monocular height estimation remains a challenging task due to the absence of inherent depth cues in single RGB images. This limitation highlights the importance of multi-task learning, which leverages complementary tasks—such as semantic segmentation—to provide structural context and enhance height prediction accuracy. Prior works that incorporate semantic priors [43, 44, 48, 12] demonstrate the synergy between tasks, motivating the design of a joint learning network. Furthermore, dual-path encoder strategies have shown promise in semantic segmentation by effectively capturing local information from high-resolution RGB inputs, especially when combined with transformer-based backbones. On the other

hand, adaptive binning approaches from monocular depth estimation have proven successful in bridging the classification-regression paradigm, and offer potential when transferred to height estimation tasks. In addition, for pixel-wise height estimation, it is observed that models often yield poor predictions around the boundaries of objects, as noted in [13]. To address this issue, a boundary-aware L1 loss is introduced, encouraging the network to better distinguish height discontinuities at object edges. When comparing backbone architectures, the VMamba backbone stands out as more efficient and accurate than alternatives like Swin Transformer and ViT, especially in extracting multi-scale hierarchical features. By integrating these findings, this thesis proposes BuildMamba, a novel multi-task learning framework that jointly addresses semantic segmentation and building height estimation. The model features a dual-path encoder combining global representations from the VMamba backbone and fine-grained local features from CNN layers, addressing the aforementioned limitations in the literature and achieving superior performance on challenging remote sensing benchmarks.

3. METHODOLOGY

This chapter introduces the proposed multi-task learning architecture, BuildMamba, developed to address key limitations observed in existing remote sensing literature —particularly underperformance in height estimation for tall buildings, lack of edge-awareness, and generalization issues across diverse urban environments. The chapter first outlines the overall architecture, which integrates a visual state-space backbone (VMamba) with a convolutional local path to capture both global semantic context and local structural features. To compensate for the absence of native attention in VMamba, a novel Mamba Attention Module (MAM) is proposed to refine spatial features stage-wise. This is followed by a multi-scale fusion using a Feature Pyramid Network (FPN), enabling the architecture to handle buildings of varying sizes effectively. The final task-specific heads are tailored separately for semantic segmentation and height estimation.

What distinguishes BuildMamba from previous works is its utilization of VMamba into the domain of remote sensing specifically for the height estimation task. Additionally, the architecture leverages a Sobel operator-based edge detection mechanism to assign higher loss weights near building boundaries, improving prediction sharpness in these regions. The model’s effectiveness is evaluated not only on standard datasets but also in terms of its generalization capability on the Huawei Building Height Estimation (BHE) dataset, which includes diverse urban morphologies. Through these innovations, BuildMamba bridges the performance-efficiency gap and provides a highly accurate and adaptable solution for large-scale urban analysis using only RGB satellite imagery.

3.1 Overview of BuildMamba Architecture

This section outlines the proposed methodology in Figure 3.1, BuildMamba, which specifically targets the challenges of accurate height estimation for tall buildings and the

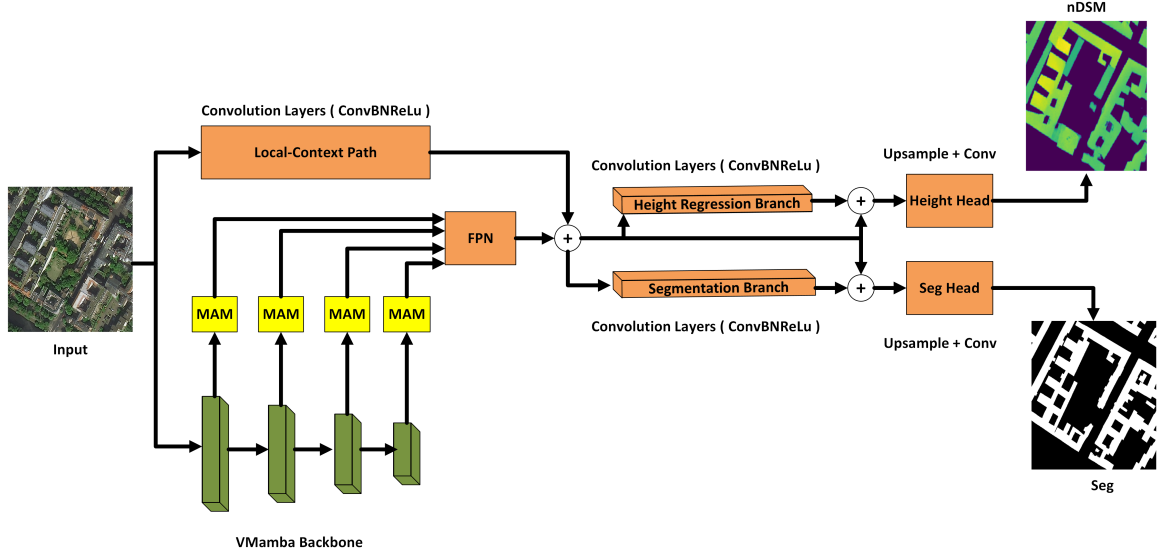


Figure 3.1: The proposed BuildMamba architecture. It consists of a dual-path encoder that fuses global features from the VMamba backbone and local features from CNN layers. The architecture includes a Mamba Attention Module (MAM), and a Feature Pyramid Network (FPN), followed by task-specific decoders for semantic segmentation and height estimation.

domination of background pixels in remote sensing imagery. The model employs a dual-path encoder architecture, similar to [3, 33] that integrates a pre-trained visual state-space backbone (VMamba) to capture global long-range dependencies, alongside CNN-based modules to model local feature context effectively. VMamba is an effective backbone compared to competitive models like [4, 41]. With its efficient and effective structure, it performs very well compared to other backbones on benchmark datasets, which is why the pre-trained version of VMamba used in this study effectively captures height and semantic cues from a single RGB image. Since VMamba does not employ an attention mechanism, a spatial-based self-attention module, termed the Mamba Attention Module (MAM), is proposed to enhance spatial feature refinement. Unlike the attention mechanisms [50, 51] which utilize parallel branches for spatial and channel-based attention modules, MAM employs only spatial attention by focusing on height and width in a parallel manner. The

aim is to focus on a point at the given location to extract height and semantic cues more effectively. As VMamba utilizes a selective-scanning mechanism, an approach conceptually similar to attention but implemented differently, integrating an explicit attention module at the output of each stage allows the model to combine the strengths of both paradigms. While both mechanisms aim to capture contextual dependencies, their complementary characteristics enhance the model's capacity for spatial reasoning. This fusion leads to improved performance in semantic segmentation and building height estimation tasks.

Furthermore, the refined feature maps from the MAM are further processed through a Feature Pyramid Network (FPN) to capture rich multi-scale contextual information. FPN's role is to capture both small and large buildings in the image, as small buildings are hard to detect for models. Following this, separate decoders are applied for building height estimation and semantic segmentation, which are used in other multi-task learning tasks [43, 44] as well. This helps both tasks optimize dedicated losses together by benefiting from semantic segmentation and height estimation tasks, where each task affects the other. Additionally, local feature maps extracted via the CNN branch are fused by summing two features into both decoder pathways to improve spatial precision [3].

There are several loss functions that are applied and tested in this thesis for building segmentation and height estimation. Since dice loss, and cross-entropy losses [3, 33, 31] are utilized in many semantic segmentation tasks, they are applied here as well, where cross-entropy loss encourages per-pixel accuracy, and dice loss is effective in the case of an imbalanced segmentation task, which is, in this case, background pixel domination over the number of building pixels. In addition to these, the Laplacian operator is applied to extract building boundaries and binary cross-entropy loss is used to distinguish building boundaries in order to help the model capture sharper boundaries. Moreover, in the case of height estimation, losses such as Huber, L1, L2, or combination of L1 and L2 are

utilized [47, 19, 12]. Each loss has different characteristics, such that L1 loss is more robust to outliers. Since outliers represent tall building structures in this case, L1 loss helps balance the imbalanced pixel distribution between background and building. On the other hand, L2 loss can penalize larger errors, but it is much more sensitive to outliers, which is not suitable for the height estimation task. Huber loss combines the robustness of L1 and the smoothness of L2 loss functions. In this study, L1 loss is chosen as the primary loss function due to its robustness to outliers and its tendency to reduce the influence of extreme errors, particularly in tall buildings. However, to mitigate its under-sensitivity to large errors in height, especially at object boundaries, a boundary-aware weighting scheme is applied to better guide the model in learning accurate height predictions for both small and tall structures.

The overall architecture of BuildMamba is illustrated in Figure 3.1. In the following subsections, detailed explanation of each component within this pipeline is provided, including the backbone (VMamba), the dual-path encoder design, the Mamba Attention Module (MAM), the Feature Pyramid Network (FPN), and the task-specific decoder heads with loss functions for each task.

3.2 VMamba

VMamba is a vision architecture that introduces an approach to sequence modeling of computer vision tasks. It is designed to replace traditional attention mechanisms with selective state-space models (SSMs). CNNs capture local detail and transformers use attention to capture global spatial dependencies from the feature maps. However, VMamba accomplishes better performance by utilizing SSMs which are designed to model long-range dependencies linearly with lower computational cost.

Vision transformers have been successful in computer vision tasks because of their capacity to handle global context with self-attention. However, their quadratic complexity

with respect to image input sizes causes a bottleneck in high resolution images. On the other hand, CNNs are efficient but incapable of modeling long range dependencies. CNNs have a local receptive field in order to capture local context of the images. VMamba aims for balance between transformers and CNNs by offering global modeling capabilities with linear computational complexity which makes it ideal for image understanding tasks that demand both accuracy and speed.

3.2.1 *Architecture Overview*

VMamba modifies the attention block used in vision transformers with Mamba blocks that are structured state-space sequence models. The main contributions allow for selective scan operations and input-dependent state updates. These innovations enable the model to capture long-range dependencies more efficiently compared to state-of-the-art transformers. As illustrated in Figure 3.2, VMamba begins with patch embedding, which splits the image into non-overlapping patches and embeds them into tokens. Then, each token is processed using Mamba blocks, which utilize state-space models to learn spatial dependencies. Mamba models sequences using recurrent state updates. Layer normalization and Multi-Layer Perceptron (MLP) layers follow each Mamba block to enhance the stability and expressiveness of the training. Like other hierarchical backbones, VMamba constructs multi-scale feature maps across consecutive stages, allowing fine-to-coarse learning.

3.2.2 *Selective State-Space Models*

At the core of VMamba lies a selective state-space model (SSSM), which offers an alternative to the attention mechanism used in classical Transformers. Instead of attending to all tokens simultaneously via self-attention, VMamba processes sequences using a recurrence-based formulation inspired by state-space models (SSMs), but with an important distinction: the recurrence is input-dependent. That is, the way previous states

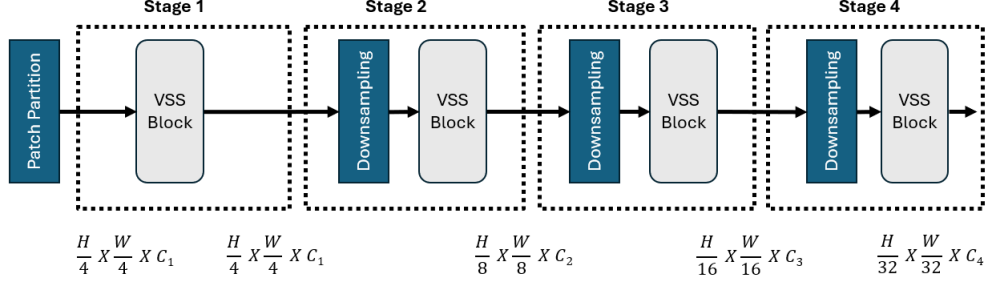


Figure 3.2: Overall architecture of VMamba [1]. The model begins with a patch embedding layer that tokenizes the input image into non-overlapping patches. These tokens are passed through multiple stages of Mamba blocks, each consisting of a state-space-based recurrent update mechanism designed to capture long-range spatial dependencies efficiently. Each stage also includes layer normalization and MLP blocks to stabilize and enhance feature learning. The architecture produces multi-scale feature maps.

influence the current state varies depending on the current input.

The recurrence is defined as:

$$h_{t+1} = A(h_t, x_t) + B(x_t) \quad (3.1)$$

where:

- $h_t \in \mathbb{R}^d$ is the hidden state at time step t ,
- $x_t \in \mathbb{R}^d$ is the input token or feature at time step t ,
- $A(h_t, x_t)$ is a learnable function that modulates the contribution of the previous hidden state h_t based on the current input x_t ,
- $B(x_t)$ is a transformation applied directly to the current input, independent of h_t .

In VMamba, both $A(\cdot)$ and $B(\cdot)$ are parameterized using lightweight convolutional and gating mechanisms. These components allow the model to selectively control information flow across time steps, making it possible to capture both local patterns and long-range dependencies, a key requirement in vision tasks. Importantly, this approach

achieves linear time complexity, offering improved scalability over traditional attention-based models while maintaining competitive performance on large-scale image recognition benchmarks.

In this study, VMamba encoder, pre-trained on ImageNet, is used as backbone of the multi-task learning network which performs semantic segmentation and height regression from satellite imagery. VMamba plays crucial role in improving the model’s performance in both areas.

3.3 Dual-Path Encoder

The proposed architecture employs a dual-path encoder designed to effectively capture both global contextual dependencies and local structural details, which are critical for accurate building segmentation and height estimation. The first path utilizes VMamba, a visual state-space model that excels at modeling long-range dependencies without relying on traditional attention mechanisms. This path is responsible for encoding global semantic context across the entire spatial domain, enabling the network to understand broad spatial relationships, such as the overall layout of urban regions or relative building structures. On the other hand, the second path consists of a convolutional neural networks that specializes in capturing fine-grained local features such as edges, corners, and texture patterns. By processing the input through both VMamba and CNN pathways in parallel, the encoder generates feature representations that are both semantically rich and spatially precise. These dual features are later fused and propagated through the network, enabling more robust and scale-aware predictions for both semantic segmentation and height regression tasks.

3.4 Mamba Attention Module (MAM)

To enhance the fusion of local and global features extracted by the dual-path encoder, a Mamba Attention Module (MAM) is integrated at the output of each stage in the VMamba encoder. Drawing inspiration from spatial attention mechanisms such as SAM [51], MAM is designed to refine feature representations by dynamically recalibrating spatial importance, allowing the network to emphasize informative regions while suppressing less relevant or noisy features, as illustrated in Figure 3.3.

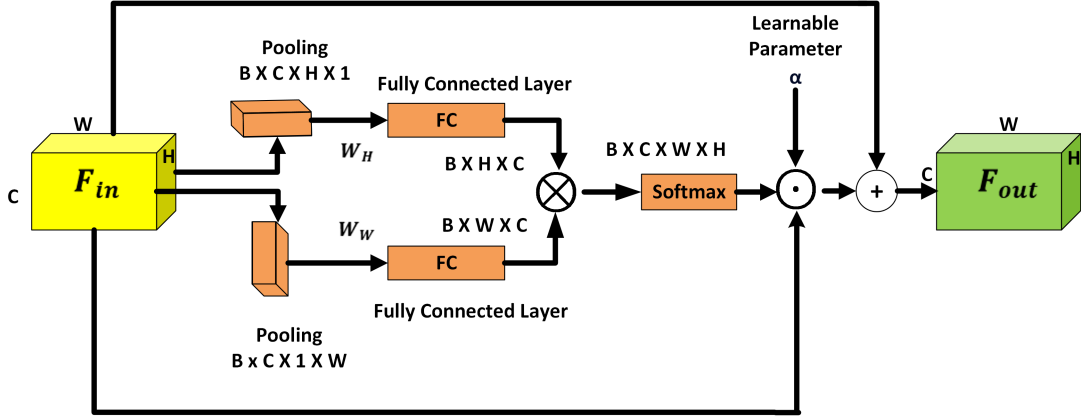


Figure 3.3: Architecture of the Mamba Attention Module (MAM). The input feature map $F_{in} \in \mathbb{R}^{C \times H \times W}$ is processed via two parallel spatial pathways: one applies adaptive average pooling along the height dimension ($H \times 1$), and the other along the width dimension ($1 \times W$). Each pooled output is passed through a fully connected layer to produce intermediate representations. These are combined via element-wise multiplication, followed by a softmax normalization to generate an attention map. This map is used to reweight the input features, which are then added back to the original feature map through a residual connection scaled by a learnable parameter α , resulting in the refined output F_{out} .

MAM processes features along two pathways by applying adaptive average pooling separately across height ($H \times 1$) and width ($1 \times W$) dimensions. This dual-pathway approach captures spatial dependencies from both horizontal and vertical perspectives. The pooled features are transformed through fully connected layers and combined to generate an attention map, which highlights significant spatial regions. The final output is

produced by applying this attention map to the input features along with a residual connection, effectively enhancing spatially relevant information. The attention map is normalized with a softmax function and subsequently applied to the original input features:

$$\mathbf{F}_{\text{out}} = \alpha \cdot \text{softmax}(\mathbf{W}_w \mathbf{P}_w \mathbf{F}_{\text{in}} \otimes \mathbf{W}_h \mathbf{P}_h \mathbf{F}_{\text{in}}) \odot \mathbf{F}_{\text{in}} + \mathbf{F}_{\text{in}}, \quad (3.2)$$

where $\mathbf{F}_{\text{out}}$ is the refined output feature map, and $\mathbf{F}_{\text{in}}$ is the input feature map. α is a learnable scalar parameter that controls the contribution of the attention mechanism. $\mathbf{P}_w$ and $\mathbf{P}_h$ are the adaptive average pooling operators along the height and width dimensions, respectively. $\mathbf{W}_w$ and $\mathbf{W}_h$ are learnable weight matrices applied via fully connected layers, responsible for transforming the pooled feature maps to capture richer dependencies before computing the attention map. The symbol $\odot$ denotes element-wise multiplication, while $\otimes$ represents the matrix multiplication.

3.5 FPN

Figure 3.4 indicates the FPN [2] architecture. It is a widely adopted architecture designed to effectively handle objects and structures at multiple spatial scales, which is particularly valuable in dense prediction tasks such as semantic segmentation and height estimation. Traditional CNNs generate hierarchical feature maps where low-level layers capture fine-grained spatial details but lack semantic richness, while high-level layers capture strong semantics but with coarse spatial resolution. FPN addresses this limitation by constructing a top-down pathway that progressively upsamples high-level semantic features and fuses them with corresponding low-level features through lateral connections. This fusion results in feature maps that are both semantically strong and spatially precise across multiple scales. The final output is a feature pyramid where each level represents features at a different resolution, enabling the network to detect and analyze both large and small objects effectively. FPN has been widely adopted in remote sensing applications, particularly in semantic segmentation tasks [3], due to its ability to capture multi-scale

building structures and benefit to the model’s performance in accurately identifying both small houses and tall skyscrapers.

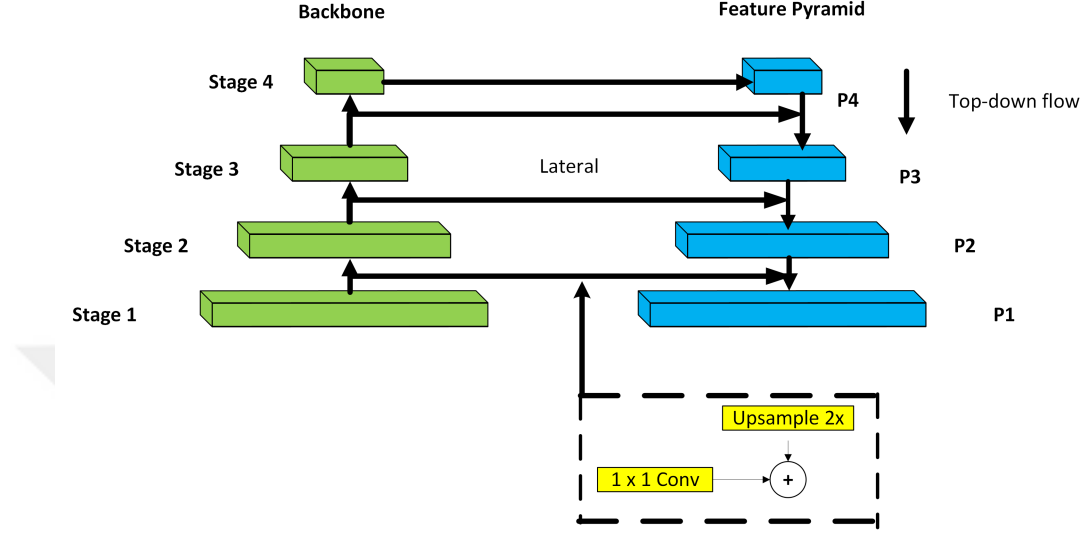


Figure 3.4: Illustration of the Feature Pyramid Network (FPN) [2] architecture. The backbone produces hierarchical feature maps from Stage 1 to Stage 4, with decreasing spatial resolution and increasing semantic richness. These features are connected to the feature pyramid via lateral 1×1 convolutions. A top-down pathway upsamples higher-level features and merges them with lower-level lateral features via element-wise addition, producing a set of multi-scale representations $\{P_1, P_2, P_3, P_4\}$. This structure enables the model to utilize both fine and coarse spatial features for downstream tasks.

3.6 Semantic Segmentation Head

Model parameters are shared up to the task-specific decoders in the multi-task learning architecture. Two separate decoders are employed to help the model distinguish between tasks and balance their contributions to the overall performance. The segmentation decoder consists of convolutional layers, each followed by ReLU activation and batch normalization at every stage. It receives an input tensor of shape $384 \times 128 \times 128$ where 384 is the channel dimension and 128×128 is the spatial resolution. Following the decoder, the segmentation head is composed of transposed convolution layers, each followed by ReLU activation and batch normalization. Two such layers are applied to

upsample the output to the original image size used in the dataset. After final transposed convolution, a final convolution layer is applied to get the final 2-class logits, which are building and background for this task. Then, those logits are processed through softmax activation to get the class scores for evaluation.

Semantic segmentation requires carefully designed loss functions to achieve optimal performance. Cross-entropy loss is used to classify each pixel into its respective class, namely building or background. Additionally, Dice loss is incorporated to address the class imbalance, particularly due to the dominance of background pixels, making it well-suited for this scenario. Finally, the Laplacian kernel [3], spatial filter used to detect regions of rapid intensity change such as edges or boundaries, is applied to extract the building contours. As a second-order derivative operator, it enhances edge features where pixel intensities change abruptly. In this work, the Laplacian kernel is used to extract the building boundaries from both the ground truth and predicted masks, and a binary cross-entropy loss is computed on these boundary maps. This encourages the model to produce sharper and more precise building outlines. The loss functions for semantic segmentation are formulated as follows:

$$\mathcal{L}_{\text{ce}}(Y, \hat{Y}) = - \sum_i Y_i \log(\hat{Y}_i), \quad (3.3)$$

$$\mathcal{L}_{\text{dice}}(Y, \hat{Y}) = 1 - \frac{2 \sum_i Y_i \hat{Y}_i}{\sum_i Y_i + \sum_i \hat{Y}_i}, \quad (3.4)$$

$$\mathcal{L}_{\text{bce}} = - \sum_{i \in E} \left[\ell_i \log(\hat{\ell}_i) + (1 - \ell_i) \log(1 - \hat{\ell}_i) \right], \quad (3.5)$$

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{dice}} + 10 \cdot \mathcal{L}_{\text{bce}}. \quad (3.6)$$

where:

- Y_i : Ground truth binary label for pixel i .
- $\hat{Y}_i$: Predicted probability for pixel i being a building.

- $\ell_i = L(Y)_i$: Laplacian edge value of ground truth at pixel i .
- $\hat{\ell}_i = L(\hat{Y})_i$: Laplacian edge value of prediction at pixel i .
- $\mathcal{L}_{\text{ce}}$: Pixel-wise cross-entropy loss.
- $\mathcal{L}_{\text{dice}}$: Dice loss to handle class imbalance.
- $\mathcal{L}_{\text{bce}}$: Binary cross-entropy loss computed over Laplacian edge map E .
- $\mathcal{L}_{\text{seg}}$: Final segmentation loss combining all components.

Laplacian Kernel L :

$$L = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

3.7 Height Estimation Head

The height estimation branch adopts a symmetrical architecture to the segmentation branch, consisting of convolutional and transpose convolutional layers with residual connections. Evaluations of model performance reveal that the model encounters challenges in accurately predicting heights at building boundaries. To address this issue, Sobel operator [13] is applied to extract building edges and assign higher weights to height predictions at the building boundaries to enhance performance on building boundaries. Furthermore, predicted and ground truth height values are normalized with logarithmic normalization since distribution of height values is long-tailed. The log-normalization is applied before feeding height values into loss function. Normalized height values are formalized as follows:

$$\hat{h}_{\text{norm}} = \frac{\log(\hat{h})}{\log(h_{\text{max}})}, \quad h_{\text{norm}} = \frac{\log(h)}{\log(h_{\text{max}})} \quad (3.7)$$

where:

- $\hat{h}$ is the predicted height value,
- h is the ground truth height value,
- $h_{\max}$ is the maximum building height in the dataset,
- $\hat{h}_{\text{norm}}$ and h_{norm} are the logarithmically normalized predicted and ground truth heights, respectively.

L1 loss [52] is employed as the primary loss function because of its robustness to outliers. In this case, tall buildings become outliers since the number of tall buildings is the lowest compared to other building types such as mid-rise and low-rise buildings. Weighted L1 loss is defined as follows:

$$\mathcal{L}_{\text{height}} = \mathcal{L}_{\text{L1}}(h_{\text{norm}}, \hat{h}_{\text{norm}}) \odot W_{\text{height}} \quad (3.8)$$

where W_{height} is the pixel-wise weighting mask and $\mathcal{L}_{\text{L1}}$ denotes the mean absolute error (L1 loss) computed element-wise. The pixel-level weights W_{height} are defined as:

$$W_{\text{height}}(i) = \begin{cases} w_{\text{boun}}, & \text{if } NS(i) > \tau, \\ w_{\text{tall}}, & \text{else if } h(i) > 40, \\ w_{\text{other}}, & \text{otherwise,} \end{cases} \quad (3.9)$$

where τ is a threshold which is set to 0.1 for edge detection, and $w_{\text{boun}}, w_{\text{tall}}, w_{\text{other}}$ are the weights for boundary, tall buildings bigger than 40 meters, and non-boundary pixels, respectively. $NS(i)$ is the normalized Sobel edge magnitude for pixel i , which is used to assign different weights to boundary and remaining pixels to emphasize boundary height prediction accuracy. $NS(i)$ is computed as follows:

$$NS(i) = \frac{\mathcal{S}(Y)(i)}{\max(\mathcal{S}(Y)) + \epsilon}, \quad (3.10)$$

where $\mathcal{S}(Y)$ denotes the Sobel-filtered ground truth map for extracting building edges, and ϵ is a small constant to avoid division by zero. After conducting several experiments, we set w_{boun} to 2, w_{tall} to 10, and w_{other} to 1.

3.8 Multi-Task Learning

In this work, we address two tasks simultaneously: building height estimation and semantic segmentation. To achieve efficient learning and parameter utilization, we employ a hard parameter sharing strategy and optimize the model using uncertainty weighting [53] for the loss functions:

$$\mathcal{L}_{\text{total}} = \frac{1}{2\sigma_{\text{height}}^2} \mathcal{L}_{\text{height}} + \frac{1}{2\sigma_{\text{seg}}^2} \mathcal{L}_{\text{seg}} + \log \sigma_{\text{height}} + \log \sigma_{\text{seg}}. \quad (3.11)$$

where σ_{height} and σ_{seg} are learnable parameters that represent the task-specific uncertainties for height estimation and segmentation, respectively. $\mathcal{L}_{\text{height}}$ is the height estimation loss and $\mathcal{L}_{\text{seg}}$ is the segmentation loss. The logarithmic terms ($\log \sigma_{\text{height}}$ and $\log \sigma_{\text{seg}}$) act as regularizers to prevent the uncertainty values from becoming too small.

To efficiently share knowledge between the two tasks, both tasks utilize the same feature extraction backbone. The shared representation serves as input to two task-specific decoders: one for building height estimation and another for segmentation.

3.9 Architectural and Loss Function Variants

In this section, the architectural and loss function variants explored in the scope of this study are presented. Specifically, different backbone architectures, decoder heads, and loss functions are evaluated to analyze their individual impact on overall model performance.

3.9.1 Backbone

Earth vision requires foundational models that are capable of understanding the spatial and semantic context of complex environments effectively. In the context of building height estimation and segmentation from satellite imagery, the choice of backbone is critical, as it determines the model’s ability to extract multi-scale, high-level features from large and high-resolution inputs. In this study, ViT [3] and VMamba [1] are utilized to understand how varying design philosophies influence performance of the model.

3.9.1.1 Vision Transformer (ViT)

ViTs [41] represent a paradigm shift in visual representation learning by treating images as sequences of patches, analogous to tokens in natural language processing. Instead of relying on convolutional operations, ViTs divide an image into fixed-size patches, flatten them, and feed them through a transformer encoder, allowing the model to capture long-range dependencies across the entire image through self-attention mechanisms.

Unlike CNNs, which inherently encode locality and spatial hierarchies through receptive fields, ViTs rely on positional encodings to maintain spatial relationships. This enables ViTs to learn more global context, which is particularly beneficial in remote sensing and Earth observation tasks where features such as buildings, roads, and terrain often span large areas and may be visually similar at different scales.

In the context of this study, BuildFormer [3] is explored as a backbone within the multi-task architecture for semantic segmentation and building height estimation since it is specifically designed for building segmentation from remote sensing images. Although Transformers demonstrates strong performance in capturing global structure and contextual information, its application to high-resolution satellite images poses several challenges. These include high memory consumption, and computational cost.

One block of the BuildFormer architecture is illustrated in Figure 3.5. In this

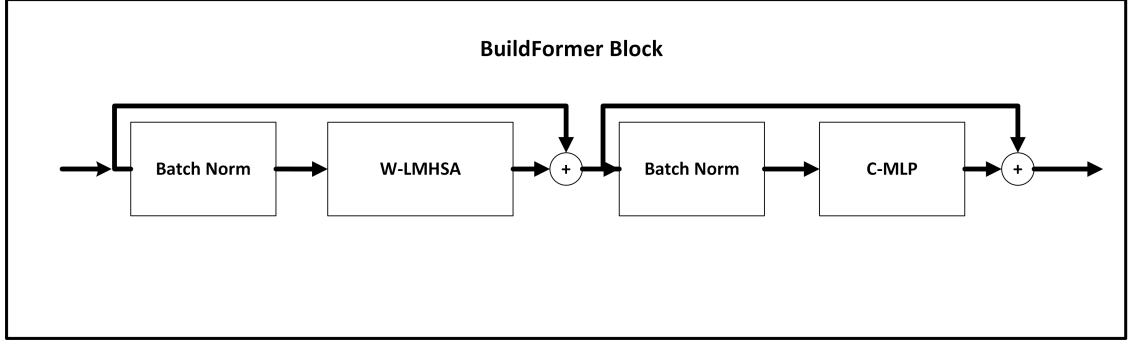


Figure 3.5: Overview of the BuildFormer [3] Block architecture. The first path begins with Batch Normalization followed by a Window-based Linear Multi-Head Self-Attention (W-LMHSA) module. This output is added back to the input via a residual connection. The second stage consists of another Batch Normalization layer followed by a Convolutional Multi-Layer Perceptron (C-MLP), responsible for enhancing channel-wise feature interactions. A final residual connection ensures gradient flow and stabilizes training.

block, the input feature map is first passed through a batch normalization layer, followed by the Window-based Linear Multi-Head Self-Attention (W-LMHSA) module. The output of this attention module is then added to the original input via a residual connection. This process is repeated in the second stage, where another batch normalization layer and a Convolutional Multi-Layer Perceptron (C-MLP) are applied. The result is again added to the intermediate output through a second residual connection, producing the final output feature map of the BuildFormer block.

3.9.1.2 Swin Transformer

Swin Transformer [4] introduces a hierarchical vision transformer architecture that addresses the scalability and efficiency limitations of the original ViT. Instead of computing global self-attention across all image patches, Swin Transformer employs a shifted window mechanism that partitions the image into non-overlapping windows, within which self-attention is computed locally. By periodically shifting these windows, Swin effectively enables cross-window connections while maintaining computational efficiency. This hierarchical design allows Swin Transformer to model both local and global

dependencies while preserving spatial hierarchies, similar to the feature pyramid structure of CNNs. As a result, it achieves a better balance between accuracy and computational cost, making it highly suitable for dense prediction tasks such as semantic segmentation, object detection, and depth estimation. In the context of this study, the Swin Transformer is examined as an alternative backbone within the proposed multi-task framework. Its ability to process high-resolution images efficiently and capture multi-scale contextual information makes it particularly advantageous for remote sensing applications, where structures like buildings and roads exhibit complex spatial arrangements at varying scales.

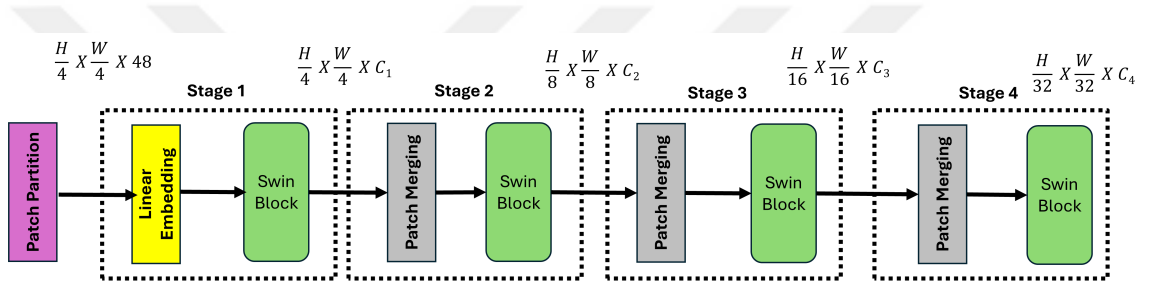


Figure 3.6: Overview of the Swin Transformer architecture [4]. The model processes the input image through a hierarchical structure consisting of four stages. Each stage includes multiple Swin Transformer blocks that operate within shifted local windows to perform self-attention efficiently. Between stages, patch merging operations downsample spatial dimensions while increasing feature depth, enabling multi-scale feature extraction. This hierarchical design allows Swin Transformer to balance local detail preservation with global context modeling

3.9.2 Heads

This section presents the different decoder heads explored for the two main tasks in this study: semantic segmentation and building height estimation. Each head is evaluated based on its architectural design and its impact on task-specific performance, particularly in complex urban environments.

3.9.2.1 AdaBins

This section presents the integration of the AdaBins module into the height estimation branch of the proposed network to improve height prediction through adaptive binning. AdaBins adopts a hybrid classification–regression approach with adaptive depth binning for more accurate continuous depth prediction. Instead of fixed bins, it dynamically generates scene-specific bin edges. Each pixel is classified into a bin, followed by regression using a weighted combination of bin centers. This adaptive strategy allows finer resolution in critical depth ranges.

To integrate AdaBins [5] module (see Figure 3.7) into the height prediction pipeline, the module is appended to the final layer of the height estimation branch. This enables the model to formulate height prediction as a discretized classification-regression task, where bin centers are adaptively computed based on the input image features. The default hyperparameters are adopted and provided for the mini-ViT backbone, ensuring lightweight yet effective bin prediction. The minimum and maximum values for bin computation are dataset-specific, with the minimum value typically close to zero and the maximum height value empirically selected based on the dataset’s observed range (refer to Table 4.2). AdaBins employs Chamfer loss [5] to enforce alignment between predicted depth distributions and their corresponding ground-truth values. This loss penalizes the distance between the predicted bin centers and the true depth points, ensuring fine-grained localization of predicted depths. However, in this case, as Chamfer loss primarily captures geometric alignment and may under-penalize absolute errors, we combine it with the L1 loss to encourage global consistency and numerical accuracy.

$$\mathcal{L}_{\text{height}} = \lambda_1 \cdot \mathcal{L}_{\text{Chamfer}} + \lambda_2 \cdot \mathcal{L}_{\text{L1}} \quad (3.12)$$

where $\mathcal{L}_{\text{Chamfer}}$ encourages geometric alignment of predicted height bins with ground-truth heights, and $\mathcal{L}_{\text{L1}}$ ensures accurate pixel-wise height estimation. In our experiments, we

empirically set the weights to $\lambda_1 = 0.001$ and $\lambda_2 = 10$.

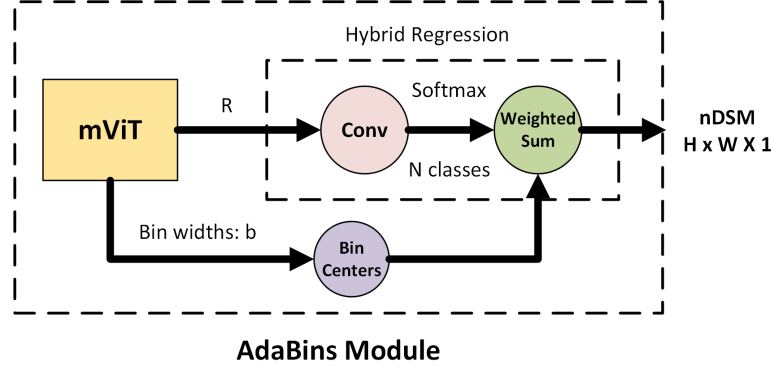


Figure 3.7: Overview of the AdaBins module [5] for hybrid regression in depth, or height estimation in this case. AdaBins module combines classification and regression in a hybrid formulation to produce continuous height outputs. The feature representation R generated by the mViT backbone is passed through a convolutional layer to predict class probabilities over N discretized height bins. A softmax is applied to obtain per-pixel class distributions. Simultaneously, adaptive bin centers are computed based on learnable bin widths b and a weighted sum is performed to regress the final height value at each pixel. This approach allows the network to discretize the continuous height space into bins.

3.9.2.2 VPDDepth

VPD decoder head [54] functions as a lightweight yet effective module designed to transform the multi-scale feature representations extracted from the diffusion-model backbone into dense prediction outputs. Specifically, the decoder integrates feature maps from several hierarchical layers of the pretrained UNet within the diffusion model, each corresponding to different spatial resolutions, to preserve both global context and fine-grained spatial details. These features are first normalized and projected through 1×1 convolutional layers to ensure channel consistency before being progressively upsampled and fused via lateral skip connections, following the principle of a FPN. Additionally, cross-attention maps derived from the diffusion process, linking textual embeddings with visual latent representations, are concatenated to enhance semantic alignment between the input prompt and image regions. The fused multi-scale features are then passed through a

task-specific prediction layer, typically a 3×3 convolution followed by normalization and activation, to produce the final output such as semantic segmentation or depth estimation. During training, only the decoder parameters are fine-tuned while the diffusion backbone remains frozen, enabling efficient adaptation of the pre-trained generative model to downstream perception tasks with minimal computational overhead. In our study, the VPD decoder head was experimented with as an alternative to the conventional CNN-based height regression head to leverage its lightweight and computationally efficient structure. The VPD decoder demonstrates strong representational cues derived from the generative diffusion model, offering a more balanced integration between semantic understanding and spatial reconstruction.

3.9.2.3 UPerNet

UPerNet [6], shown in Figure 3.8, serves as a hierarchical feature aggregation and prediction module built upon the principles of the FPN and the Pyramid Pooling Module (PPM). This design aims to effectively integrate both global contextual and fine-grained spatial information extracted from the multi-level feature maps of the backbone network. Specifically, the decoder first applies a PPM to the deepest feature map to enrich it with global priors through multi-scale pooling operations. The PPM output is then propagated into the FPN’s top-down pathway, where higher-level semantic features are progressively upsampled and fused with corresponding lower-level features via lateral 1×1 convolutional connections. Each fusion stage generates a new set of feature maps $\{P2, P3, P4, P5\}$, which preserve the hierarchical semantics at different spatial resolutions. After normalization and ReLU activation, all feature maps are bilinearly upsampled to the highest spatial resolution and concatenated to form a unified multi-scale representation. A final 3×3 convolutional layer is then applied to refine and project these fused features into the desired output space, such as segmentation masks. UPerNet is optimized

jointly with the backbone in a multi-task manner, allowing it to capture complementary cues across hierarchical levels while maintaining computational efficiency. This structured fusion strategy enables UPerNet to produce semantically rich and spatially consistent dense predictions across complex urban scenes. Moreover, recent models [4, 1] adopt UPerNet as the segmentation head; therefore, it was chosen for experimentation in this study to enhance the model’s semantic understanding.

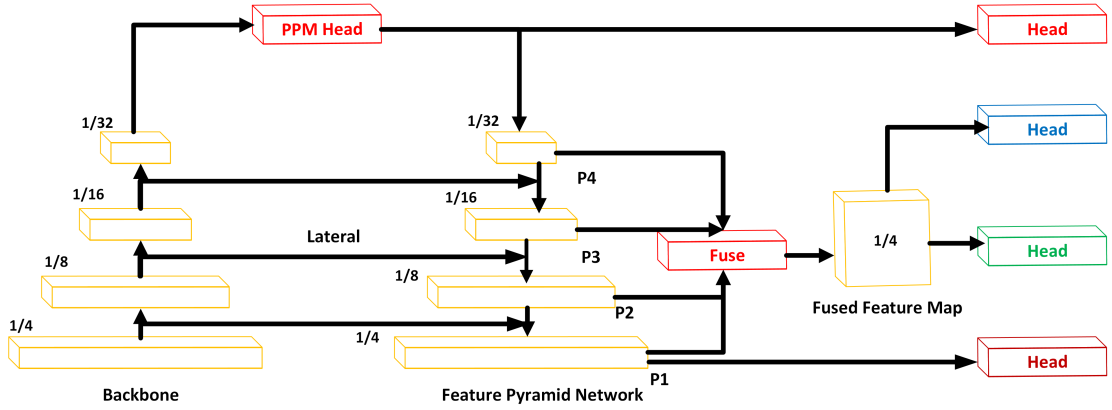


Figure 3.8: Overview of the UPerNet architecture [6] for dense prediction tasks, adapted here for semantic segmentation. UPerNet leverages a unified framework built upon a Feature Pyramid Network to integrate multi-scale features extracted from the backbone network. Specifically, features from different stages (e.g., $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, $\frac{1}{32}$) are first aligned through lateral connections and then fused to obtain a semantically enriched representation. PPM is applied to the deepest feature to capture global context, which is then merged with the upsampled features. Finally, the aggregated feature map is fed into a lightweight decoder head for semantic segmentation. This architecture enables effective fusion of both local spatial details and global contextual information across scales.

3.9.2.4 Atrous Spatial Pyramid Pooling (ASPP)

ASPP [7] head, shown in Figure 3.9, is utilized to capture multi-scale contextual information by applying parallel atrous convolutions with varying dilation rates. This structure allows the model to enlarge its receptive field without increasing the number of parameters or losing spatial resolution, making it particularly suitable for dense prediction tasks such as semantic segmentation. Specifically, the ASPP module processes the

highest-level feature map obtained from the backbone through several parallel branches: a 1×1 convolution for local feature extraction, three 3×3 atrous convolutions with different dilation rates to capture context at multiple scales, and a global average pooling branch that encodes global scene information. The outputs of these branches are concatenated along the channel dimension and passed through another 1×1 convolution to fuse the multi-scale features into a unified representation. Batch normalization and ReLU activation are subsequently applied to enhance stability and non-linearity. Finally, a task-specific prediction layer produces the dense output, such as segmentation maps. The ASPP head is optimized together with the rest of the network in order to capture large-scale context understanding, which is essential for modeling complex urban structures.

3.9.3 Loss Functions

This section outlines the loss functions explored for the height estimation task. Since pixel-wise building height prediction is a regression problem, selecting an appropriate loss function plays a key role in balancing error sensitivity and robustness to outliers.

3.9.3.1 Huber Loss

Huber Loss [55], combines the best properties of L1 and L2 losses. It behaves like L2 loss for small errors and like L1 loss for large errors which makes it less sensitive to outliers. This is often advantageous in height estimation tasks, where extreme height values can otherwise dominate the loss function and negatively affect the learning process.

The Huber loss is defined as:

$$\mathcal{L}_{\text{Huber}}(y_i, \hat{y}_i) = \begin{cases} \frac{1}{2}(y_i - \hat{y}_i)^2, & \text{for } |y_i - \hat{y}_i| \leq \delta, \\ \delta (|y_i - \hat{y}_i| - \frac{1}{2}\delta), & \text{otherwise.} \end{cases} \quad (3.13)$$

where y_i is the ground truth, $\hat{y}_i$ is the predicted value, and δ is a tunable threshold that determines the transition point between L2-like and L1-like behavior.

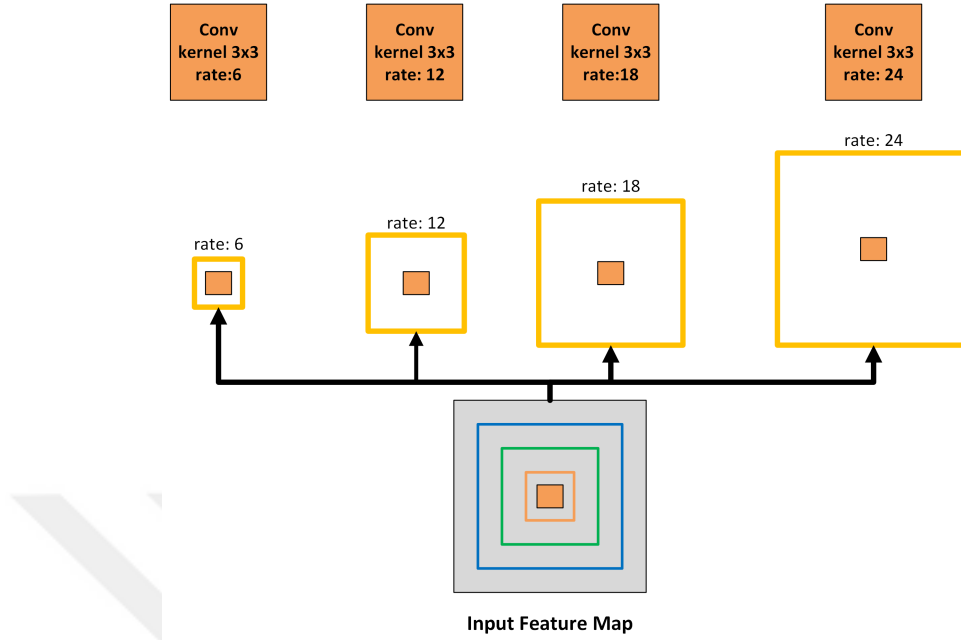


Figure 3.9: Overview of the Atrous Spatial Pyramid Pooling (ASPP) module [7], which is used as an alternative segmentation head to analyze the performance of the proposed model. ASPP module enhances the receptive field by applying multiple parallel atrous convolutions with different dilation rates to the same input feature map. Each branch captures contextual information at a distinct scale, allowing the network to effectively integrate both local details and long-range dependencies. The outputs from all branches are concatenated and fused through a 1×1 convolution followed by batch normalization and activation to produce the final feature representation. This structure enables multi-scale context aggregation and ability to segment complex urban structures in satellite imagery.

3.9.3.2 L1 Loss

The L1 Loss [52], also known as MAE, measures the average absolute difference between predicted and ground truth values. It is robust to outliers compared to L2 loss but lacks the smooth gradient properties that Huber Loss offers. Despite its simplicity, L1 Loss is effective for height estimation tasks where preserving fine-grained differences is more important than penalizing large deviations.

The L1 loss is defined as:

$$\mathcal{L}_{L1}(y_i, \hat{y}_i) = |y_i - \hat{y}_i|. \quad (3.14)$$

where y_i and $\hat{y}_i$ denote the ground truth and predicted height values at pixel i , respectively.



4. EXPERIMENTS AND RESULTS

This chapter presents the experimental setup, implementation details, and evaluation results of the proposed BuildMamba architecture. First, the datasets used throughout the study are introduced, including DFC19, DFC23, and the Huawei Building Height Estimation (BHE) dataset. Next, the implementation details—such as training configurations, hyperparameters, and pre-processing steps—are outlined to ensure reproducibility. Following this, the results section reports the performance of BuildMamba on both building segmentation and height estimation tasks, including comparisons with existing baselines. Finally, a comprehensive discussion and ablation study are conducted to analyze the contribution of each architectural component, including the VMamba backbone, Mamba Attention Module (MAM), and Sobel-based boundary-aware loss weighting. This section aims to offer insights into how each module enhances the model’s accuracy, robustness, and generalization ability.

4.1 Datasets

In this section, the DFC19, DFC23, and Huawei BHE datasets used in this study are introduced. The characteristics of each dataset are detailed, including spatial resolution, annotation types, and urban diversity. Additionally, the dataset splits into training, validation, and test sets are explained along with the pre-processing steps applied prior to model training.

4.1.1 DFC19

The DFC19 dataset [11] is a large-scale remote sensing benchmark that provides single-view satellite imagery along with ground truth geometric and semantic annotations for urban regions in Jacksonville, Florida, and Omaha, Nebraska. The dataset spans approximately 100 km² and includes imagery collected between 2014 and 2016 at a ground



Figure 4.1: *Visual Comparison of Images in Each Dataset.*

sampling distance (GSD) of 1.3 meters. It comprises 2,783 triplets of RGB images, normalized digital surface models representing above-ground-level height annotations, and semantic maps that preserve detailed building footprints, all at a resolution of 1024×1024 pixels.

The DFC19 dataset originally includes multi-class semantic labels and nDSMs designed for general-purpose 3D reconstruction tasks. To tailor the dataset specifically for building footprint extraction and height estimation, we applied a preprocessing pipeline to isolate building labels from the semantic maps. To prepare the dataset for training, NaN values in the nDSMs are imputed using nearest neighbor interpolation to ensure completeness. A binarization step is then applied to the semantic maps to isolate building footprints, simplifying the task to pure building extraction and height estimation. For speed and the required image size for the model, a random cropping strategy is adopted, generating 512×512 pixel crops from the original tiles. The number of crops per image is computed dynamically to meet a target of 20,000 total crops, which are subsequently split into 14000 training, 2000 validation, and 4000 testing subsets, while ensuring that each subset contains crops from different groups of images.

Table 4.1: *Dataset statistics for the DFC19 dataset. The table summarizes the number of images and buildings, as well as key numerical statistics relevant to building height analysis and model development.*

Statistic	Value
Total Images	20000
Total Buildings	470737
Buildings Higher Than 40 m	1893
Mean Building Height (m)	10.98
Std. Building Height (m)	12.98
Max Building Height (m)	186.50

4.1.2 DFC23

The DFC23 (Data Fusion Contest 2023) Track-2 dataset [56], released by IEEE GRSS, provides multi-source remote sensing data aimed at building height estimation and instance segmentation tasks. While the original dataset includes synthetic aperture radar (SAR) imagery, this study focuses exclusively on using RGB satellite images to predict both height maps and semantic labels. For compatibility with the semantic segmentation task, the instance segmentation labels are converted into semantic masks. The training dataset is split into 1,417 training images, 178 validation images, and 178 test images, as reference labels for the official validation and test sets are unavailable.

4.1.3 Huawei BHE

The Huawei BHE dataset is a rich and diverse collection of imagery sourced from 14 distinct cities, each selected to capture a variety of urban characteristics. These cities differ notably in factors such as sun angle, daytime lighting conditions, building height distributions, and building orientation. Some cities feature buildings captured from a top-down perspective, while others present angled or oblique views, increasing the complexity of accurately interpreting building heights from imagery. All images in the dataset

Table 4.2: *Dataset statistics for the DFC23 dataset. The table summarizes the number of images and buildings, as well as key numerical statistics relevant to building height analysis and model development.*

Statistic	Value
Total Images	1773
Total Buildings	35865
Buildings Higher Than 40 m	1435
Mean Building Height (m)	14.82
Std. Building Height (m)	12.56
Max Building Height (m)	183.00

are acquired at a level 18 resolution, ensuring high spatial detail. For evaluation and benchmarking purposes, City3 is designated as the validation set and City6 as the test set, enabling robust testing of model generalization across varying urban layouts and environmental conditions. This dataset presents a challenging yet valuable benchmark for height estimation and urban structure understanding in remote sensing applications.

Table 4.3: *Dataset statistics for the Huawei BHE dataset. The table summarizes the number of images and buildings, as well as key numerical statistics relevant to building height analysis and model development.*

Statistic	Value
Total Images	8247
Total Buildings	65963
Buildings Higher Than 40 m	5816
Mean Building Height (m)	19.24
Std. Building Height (m)	21.51
Max Building Height (m)	568.00

4.1.4 Dataset Comparison

The datasets used for building segmentation and height estimation, Huawei BHE, DFC23, and DFC19 as illustrated in Figure 4.1, exhibit substantial variability in dataset

size, complexity and building height distributions, all of which pose unique challenges for model generalization and accuracy. Table 4.3 for the Huawei BHE dataset indicates that this dataset is the most diverse in terms of height variation, featuring 65,963 buildings across 8,247 images, with a mean building height of 19.24 meters, a standard deviation of 21.51, and a maximum height reaching 568 meters. The high number of tall buildings (5,816 exceeding 40 meters) indicates a strong presence of high-rise structures, demanding models capable of distinguishing fine-grained elevation differences over a broad dynamic range. On the other hand, strong presence of high-rise buildings makes this dataset challenging.

In contrast, the DFC23 dataset, as shown in Table 4.2, is comparatively smaller, with 1,773 images and 35,865 buildings, but still includes 1,435 buildings over 40 meters, showing moderate height diversity. The mean height here drops to 14.82 meters, with a standard deviation of 12.56. Meanwhile, DFC19, the largest in terms of image and building count (20,000 images and 470,737 buildings), has the lowest average building height at 10.98 meters and a standard deviation of 12.98, despite its massive scale. Notably, Table 4.1 shows that only 1,893 buildings exceed 40 meters, highlighting the dominance of low-rise structures, which could bias models toward underestimating taller buildings.

These imbalances, particularly the skewed distribution toward low-rise structures in DFC19 and DFC23 introduce significant learning challenges for traditional height estimation networks. Regression-based models often fail to represent tall buildings accurately, as the networks tend to minimize loss across the majority (low-height) samples, leading to systematic underestimation of high-rise structures. To address this, BuildMamba incorporates a multi-task framework designed to capture both global dependencies and fine local details across spatial regions. The use of semantic priors through

segmentation further aids in distinguishing structurally meaningful regions (like building footprints) from background, enabling the model to maintain height fidelity in high-rise zones, even when such structures are underrepresented. In addition, the spatial attention-based module, MAM, enhances the model’s ability to extract semantic information from satellite imagery, while the Boundary-Aware Loss mitigates height estimation errors around building edges.

Ultimately, the dataset statistics emphasize the importance of designing height estimation models that are robust to both data imbalance and high variance in building height — a challenge that BuildMamba directly addresses through its architecture and training strategy.

4.2 Implementation Details

The proposed model was trained on the DFC19, DFC23 track 2, and Huawei BHE datasets for evaluation. The training setup was implemented on a NVIDIA GeForce RTX 2080 Ti. The initial learning rate was set to $5e-4$ for the entire network, while a lower learning rate of $5e-5$ was applied to the pre-trained VMamba backbone for fine-tuning. A weight decay of 0.0025 was employed to regularize the model and prevent overfitting. The training process was configured to run for a maximum of 465 epochs for DFC23, 225 epoch for DFC19, and 15 epoch for Huawei BHE dataset which were determined to be sufficient for convergence. The batch size was set to 8. A Cosine Annealing Warm Restarts scheduler was used. The scheduler restarted every 15 epochs, with the restart interval progressively doubling after each restart.

4.3 Metrics

In this section, the model performance evaluation metrics are explained. The evaluation is performed using three widely adopted metrics: Intersection over Union (IoU)

and F1 Score for the semantic segmentation task, and Root Mean Square Error (RMSE) along with its variants tailored for building height estimation.

To evaluate the performance of the proposed model across both semantic segmentation and height estimation tasks, a set of complementary metrics is employed. For segmentation, IoU measures the average overlap between predicted and ground truth masks across classes, serving as a robust indicator of segmentation accuracy. Alongside this, the F1 Score captures the balance between precision and recall. It gives a balanced view of a model's ability to detect true positives while avoiding false positives and false negatives.

For evaluating height estimation, several forms of RMSE are used. The standard RMSE computes the average squared error between predicted and actual height values over all pixels, including background. To focus specifically on building areas, RMSE-B evaluates the error only on pixels corresponding to building footprints, thus providing a more meaningful assessment of elevation accuracy. Additionally, the error is stratified into three height-based variants: RMSE-LOW, RMSE-MID, and RMSE-HIGH, which measure model performance on low-rise (1—15 m), mid-rise (15—40 m), and high-rise (≥ 40 m) buildings, respectively. This stratification helps analyze the model's ability to generalize across different scales, highlighting whether it can handle rare, tall structures or if it tends to regress toward the mean. Collectively, these metrics provide a comprehensive evaluation of the model's ability to detect buildings and estimate their heights accurately across a range of urban scenarios.

Beyond RMSE-based measures, Delta Accuracy Metrics ($\delta_1, \delta_2, \delta_3$) are employed to evaluate the ratio-based accuracy between predicted and ground truth heights. These metrics quantify the percentage of pixels whose predicted heights fall within specific multiplicative thresholds (1.25, 1.25^2 , and 1.25^3) of the true height values, offering insight into how closely the predictions align with actual measurements. Finally, the Coefficient of Determination (R^2) is used to assess the correlation between predicted and ground truth

heights, indicating how well the model explains variance in the data. High delta accuracies and an R^2 value approaching 1 reflect the model's ability to produce consistent and physically meaningful height estimations.

The mathematical definitions are as follows:

- **Intersection over Union (IoU):**

$$\text{IoU} = \frac{TP}{TP + FP + FN} \quad (4.1)$$

where TP , FP , and FN denote the number of true positives, false positives, and false negatives for the *building* class only. The background class is not included in the IoU computation.

- **F1 Score:**

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN} \quad (4.2)$$

where TP , FP , and FN are the total number of true positives, false positives, and false negatives for building class.

- **Root Mean Square Error (RMSE):**

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (4.3)$$

where N is the total number of evaluated pixels, y_i is the ground truth height, and $\hat{y}_i$ is the predicted height value.

- **RMSE over Building Pixels (RMSE-B):**

$$\text{RMSE-B} = \sqrt{\frac{1}{|B|} \sum_{i \in B} (y_i - \hat{y}_i)^2}, \quad B = \{i \mid \text{pixel } i \text{ belongs to a building}\} \quad (4.4)$$

- **RMSE Variants by Height Range:**

$$\text{RMSE-HIGH} = \sqrt{\frac{1}{|H|} \sum_{i \in H} (y_i - \hat{y}_i)^2}, \quad H = \{i \mid y_i \geq 40\} \quad (4.5)$$

$$\text{RMSE-MID} = \sqrt{\frac{1}{|M|} \sum_{i \in M} (y_i - \hat{y}_i)^2}, \quad M = \{i \mid 15 \leq y_i < 40\} \quad (4.6)$$

$$\text{RMSE-LOW} = \sqrt{\frac{1}{|L|} \sum_{i \in L} (y_i - \hat{y}_i)^2}, \quad L = \{i \mid 1 \leq y_i < 15\} \quad (4.7)$$

where B , H , M , and L represent the sets of pixels corresponding to all buildings, high-rise buildings, mid-rise buildings, and low-rise buildings, respectively.

- **Delta Accuracy Metrics:**

$$\delta_1 = \frac{1}{N} \sum_{i=1}^N \mathbb{I} \left(\max \left(\frac{y_i}{\hat{y}_i}, \frac{\hat{y}_i}{y_i} \right) < 1.25 \right) \quad (4.8)$$

$$\delta_2 = \frac{1}{N} \sum_{i=1}^N \mathbb{I} \left(\max \left(\frac{y_i}{\hat{y}_i}, \frac{\hat{y}_i}{y_i} \right) < 1.25^2 \right) \quad (4.9)$$

$$\delta_3 = \frac{1}{N} \sum_{i=1}^N \mathbb{I} \left(\max \left(\frac{y_i}{\hat{y}_i}, \frac{\hat{y}_i}{y_i} \right) < 1.25^3 \right) \quad (4.10)$$

where $\mathbb{I}(\cdot)$ is the indicator function, which returns 1 if the condition is true and 0 otherwise. These metrics indicate the percentage of pixels where the predicted height is within a threshold factor of the ground truth.

- **Coefficient of Determination (R^2 Score):**

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (4.11)$$

where $\bar{y}$ is the mean of the ground truth height values. The R^2 score measures how well the predicted values approximate the actual data, with 1 indicating perfect prediction.

4.4 Results

Semantic segmentation performance of the proposed BuildMamba is evaluated against six different state-of-the-art models—BuildFormer [3], DSMNet [44], ST-UNet [33], PFSegNets [34], UANet [35], FarSeg++ [32] across three benchmark datasets, namely

DFC19, DFC23, and the Huawei BHE dataset. Performance is measured using two standard metrics IoU and F1 score, providing a comprehensive view of segmentation accuracy and consistency.

On the DFC19 dataset, Table 4.4 indicates that BuildMamba significantly outperforms all baseline models with an IoU of 0.903 and an F1 score of 0.949. It surpasses the next best model, UANet, which has the closest scores to BuildMamba with an IoU of 0.890 and an F1 of 0.941. PFSegNets achieves a 0.853 IoU and 0.904 F1, while ST-UNet and FarSeg++ also demonstrates strong results but falls short of BuildMamba’s performance by a notable margin. DSMNet shows the lowest scores on both metrics, struggling to capture semantic information from the DFC19 dataset. This indicates BuildMamba’s superior capability in extracting meaningful semantic features from complex urban scenes compared to the other models.

Table 4.4: *Semantic Segmentation Performance Across Models on DFC19 Dataset*

Models	IoU	F1
BuildFormer	0.790	0.883
DSMNet	0.450	0.627
ST-UNet	0.845	0.892
PFSegNets	0.853	0.904
FarSeg++	0.814	0.897
UANet	0.890	0.941
BuildMamba	0.903	0.949

Similarly, on the DFC23 dataset, Table 4.5 shows that BuildMamba again achieves the highest performance with an IoU of 0.927 and an F1 score of 0.962, maintaining a consistent lead over competitors. Its closest competitor, UANet, also performs strongly, achieving an IoU of 0.919 and an F1 score of 0.957. Notably, while PFSegNets (0.863 IoU), DSMNet (0.849 IoU), and FarSeg++ (0.832 IoU) exhibit other competitive results,

BuildMamba’s refined semantic reasoning allows it to maintain top accuracy and robustness across variations in data distribution between DFC19 and DFC23.

Table 4.5: *Semantic Segmentation Performance Across Models on DFC23 Dataset*

Models	IoU	F1
BuildFormer	0.782	0.878
DSMNet	0.849	0.920
ST-UNet	0.824	0.903
PFSegNets	0.863	0.925
FarSeg++	0.832	0.908
UANet	0.919	0.957
BuildMamba	0.927	0.962

The Huawei BHE dataset, characterized by diverse city geometries and challenging viewpoints, reveals a stark performance drop across all models, highlighting the complexity of this dataset. Despite the overall drop, BuildMamba retains its leading position with a IoU of 0.586 and an F1 score of 0.727. UANet follows as the next best model, achieving an IoU of 0.530 and an F1 score of 0.692. Other models like ST-UNet (0.518 IoU) and FarSeg++ (0.519 IoU) show nearly identical moderate performance, while PFSegNets scores slightly lower (0.492 IoU). In contrast, models like BuildFormer and DSMNet record the lowest identical scores (IoU: 0.423, F1: 0.595). This result underscores BuildMamba’s superior generalization capabilities under spatial and geometric variability.

Overall, these results demonstrate that BuildMamba consistently achieves state-of-the-art segmentation performance across standard and challenging datasets, affirming its robustness, adaptability, and enhanced contextual understanding for urban scene parsing while indicating promising results in generalization capacity.

The visual comparisons presented in Figures 4.2, 4.3, and 4.4 offer a comprehensive evaluation of semantic segmentation performance across the DFC19, Huawei BHE

Table 4.6: *Semantic Segmentation Performance Across Models on Huawei BHE Dataset*

Models	IoU	F1
BuildFormer	0.423	0.595
DSMNet	0.423	0.595
ST-UNet	0.518	0.683
PFSegNets	0.492	0.66
FarSeg++	0.519	0.683
UANet	0.530	0.692
BuildMamba	0.586	0.727

datasets, and DFC23 respectively. Each dataset has unique characteristics that significantly influence the performance of the models.

Figure 4.2 illustrates the semantic segmentation results of various models on a sample from the DFC19 dataset. The ground truth clearly delineates building structures, providing a reliable reference for comparison. Among the models, BuildMamba demonstrates the most accurate segmentation, closely matching the building shapes and edges in the ground truth, particularly for the large central building and smaller structures along the road. DSMNet, on the other hand, produces oversmoothed predictions and fails to separate individual building instances, resulting in merged structures and reduced geometric clarity. BuildFormer and ST-UNet show improved performance over DSMNet, producing cleaner segmentations, but they still miss some finer building edges compared to BuildMamba. UANet, despite leveraging uncertainty-aware attention mechanisms, suffers from noticeable errors along the fragmented parts of the building boundaries, especially in curved or complex regions, as seen in the lower-right structure. Overall, this comparison highlights BuildMamba’s superior ability to preserve structural integrity and reduce noise in building footprint extraction on the DFC19 dataset.

Figure 4.3 presents the semantic segmentation outputs of several baseline models alongside BuildMamba model on the Huawei BHE dataset. The ground truth reveals

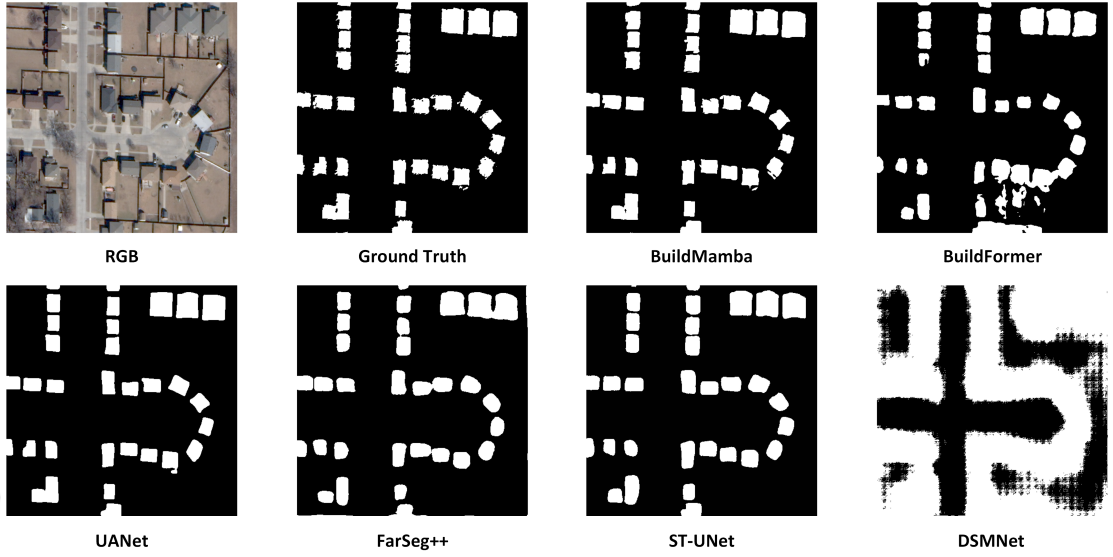


Figure 4.2: Visualization of predicted building segmentations on the DFC19 dataset, comparing results across competing models.

densely packed, elongated building structures that require precise boundary delineation for successful segmentation. Among all methods, BuildMamba produces the cleanest and most structurally faithful masks, accurately capturing the narrow buildings and preserving their elongated shapes with minimal noise. In contrast, DSMNet struggles significantly, introducing heavy artifacts and over-segmentation across the scene, leading to blurred and inconsistent building footprints. BuildFormer, ST-UNet, UANet, FarSeg++ achieve more coherent predictions than DSMNet, yet they still exhibit boundary smoothing, thinning of the building masks and a higher number of false positives/negatives. These qualitative observations align with the quantitative results in Table 4.6, where BuildMamba attains the highest IoU (0.586) and F1 score (0.727), outperforming all other baselines by a clear margin. Overall, the results demonstrate that BuildMamba excels at extracting accurate building footprints in challenging high-density scenes, maintaining structural integrity while effectively suppressing noise.

Figure 4.4 shows the semantic segmentation outputs of several baseline models

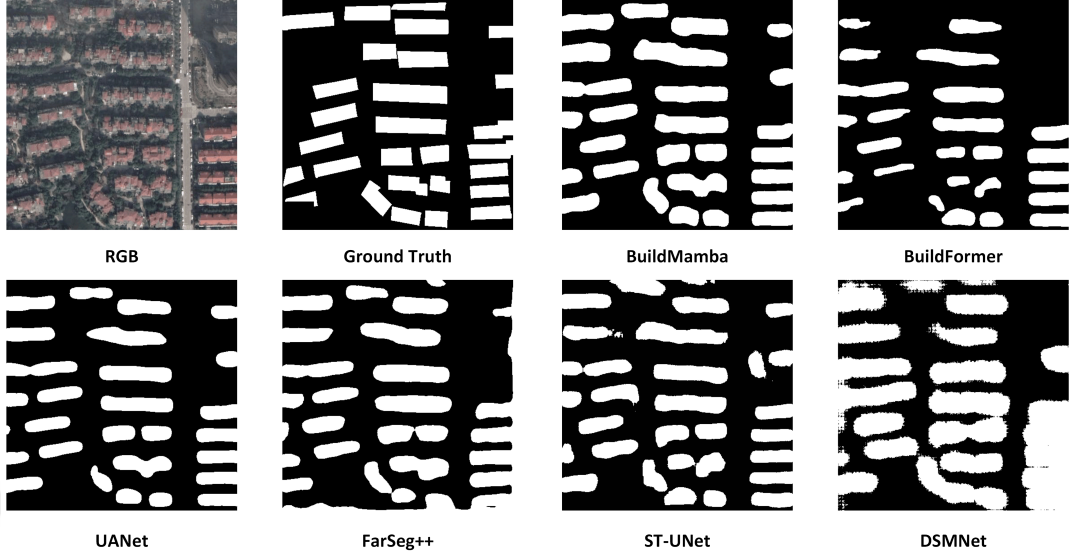


Figure 4.3: Visualization of predicted building segmentations on the Huawei BHE dataset, comparing results across competing models.

on a representative sample from the DFC23 dataset. The ground truth contains densely arranged residential buildings with irregular shapes, making precise boundary extraction essential for accurate segmentation. Among all models, BuildMamba delivers the closest match to the ground truth, successfully capturing the detailed contours of individual buildings while preserving separation between closely spaced structures. Competing models exhibit varying levels of degradation: BuildFormer produces overly smoothed masks that blur building edges, while ST-UNet and FarSeg++ tend to merge adjacent buildings or miss fine structural details. DSMNet, although more stable than on other datasets, still shows noticeable artifacts around the rooftops. UANet achieves relatively strong performance, yet it fails to maintain the same level of consistency in boundary sharpness and object completeness as BuildMamba. These qualitative findings are reflected in the quantitative results in Table 4.5, where BuildMamba attains the highest IoU (0.927) and F1 score (0.962), outperforming all other baselines by a significant margin. Overall, the results highlight BuildMamba’s superior reliability in extracting accurate and well-defined

building footprints on the DFC23 dataset, even in challenging densely built suburban environments.

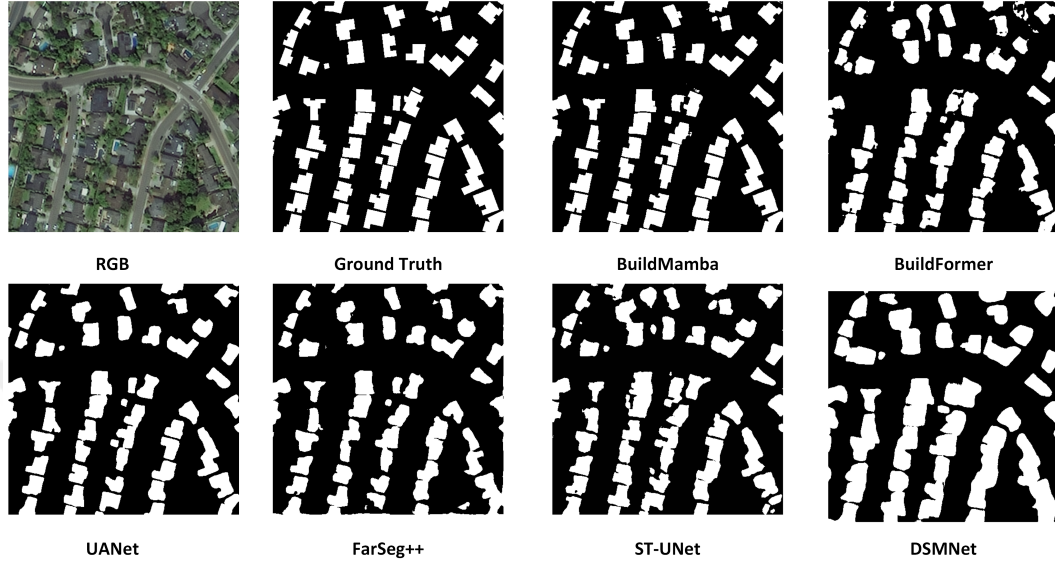


Figure 4.4: Visualization of predicted building segmentations on the DFC23 dataset, comparing results across competing models.

Building height estimation results evaluated to compare the performance of four models: LUMNet [13], DSMNet [44], IM2HEIGHT [19], IM2ELEVATION [20], and the proposed BuildMamba using three benchmark datasets: DFC19, DFC23, and Huawei BHE. Performance is assessed using Root Mean Square Error (RMSE) across all pixels and broken down into four categories: RMSE (overall), RMSE-B (buildings only), and per-building height classes—HIGH, MID, and LOW. These results highlight each model’s effectiveness in accurately predicting structural height across a variety of urban environments and elevation scales.

Table 4.7 summarizes the quantitative performance of several height estimation networks on the DFC19 dataset. The results reveal substantial variation among methods in their ability to estimate building heights across different elevation ranges. DSMNet achieves moderate accuracy, with an overall RMSE of 3.88 m and an RMSE-B of

6.65 m. However, its error rises sharply for tall structures (RMSE-H: 49.97 m), indicating difficulty in representing high-rise buildings despite relatively stable low- and mid-rise predictions. IM2HEIGHT and IM2ELEVATION both exhibit higher overall errors (4.90 m and 4.95 m, respectively) and even larger discrepancies in tall-building estimation (RMSE-H exceeding 56 m). Their low R^2 values (-0.219 and -0.262) suggest that the predicted height maps poorly correlate with ground truth, likely due to limited spatial context and over-smoothing effects during regression. LUMNet delivers a substantial improvement over these baselines, achieving an overall RMSE of 1.04 m and an R^2 of 0.92. Its balanced errors across building categories (RMSE-H: 15.56 m, RMSE-M: 6.91 m, RMSE-L: 1.62 m) and high δ -accuracy scores ($\delta_1 = 0.94$, $\delta_2 = 0.96$, $\delta_3 = 0.97$) indicate that it captures both local details and tall-structure variability more effectively. Finally, BuildMamba achieves comparable overall precision (RMSE: 1.12 m, RMSE-B: 2.12 m) and competitive performance across all height categories (RMSE-H: 12.97 m, RMSE-M: 7.98 m, RMSE-L: 1.77 m), accompanied by strong consistency metrics ($\delta_1 = 0.92$, $\delta_2 = 0.932$, $\delta_3 = 0.937$, $R^2 = 0.88$). While LUMNet achieves slightly better performance, BuildMamba maintains similar accuracy with a simpler and more stable multi-task formulation, demonstrating reliable estimation across diverse urban morphologies and effective handling of both low- and high-rise structures. Also, the superior performance of LUMNet can be attributed to its use of additional input data, specifically ground truth building masks alongside RGB imagery. In contrast, our objective is to estimate building heights solely from single RGB images. Therefore, it is reasonable that LUMNet surpasses our method in certain metrics, given its access to richer input information.

In Table 4.8, BuildMamba leads all methods with an overall RMSE of 1.95 m and a building-only RMSE of 2.67 m, outperforming DSMNet (3.39 m / 6.01 m) and

Table 4.7: *Building Height Estimation Performance on DFC19 Dataset.*

Models	RMSE	RMSE-B	RMSE-H	RMSE-M	RMSE-L	δ_1	δ_2	δ_3	R^2
DSMNet	3.883	6.654	49.967	14.480	4.091	0.443	0.476	0.500	0.200
IM2HEIGHT	4.899	9.638	56.278	18.342	6.815	0.201	0.203	0.205	-0.219
IM2ELEVATION	4.952	9.963	56.772	18.809	7.181	0.774	0.774	0.774	-0.262
LUMNet	1.043	2.034	15.560	6.913	1.622	0.942	0.963	0.971	0.922
BuildMamba	1.124	2.123	12.971	7.983	1.774	0.920	0.932	0.937	0.880

IM2HEIGHT (7.21 m / 13.89 m). It shows remarkable improvement in predicting high-rise buildings (13.93 m) compared to the closest alternatives (> 22.9 m), while maintaining low errors for mid-rise (4.00 m) and low-rise (2.74 m) categories. LUMNet, on the other hand, performs competitively with an overall RMSE of 3.62 m and an R^2 score of 0.644, showing balanced accuracy across different height levels (RMSE-H: 22.93 m, RMSE-M: 7.81 m, RMSE-L: 4.85 m). However, its performance remains limited compared to BuildMamba, particularly in high-rise regions where error accumulation is more pronounced. These results affirm BuildMamba’s generalization capabilities in more varied urban layouts and height distributions.

Table 4.8: *Building Height Estimation Performance on DFC23 Dataset.*

Models	RMSE	RMSE-B	RMSE-H	RMSE-M	RMSE-L	δ_1	δ_2	δ_3	R^2
DSMNet	3.389	6.013	24.485	8.097	4.939	0.746	0.805	0.849	-0.228
IM2HEIGHT	7.210	13.890	49.803	21.794	8.502	0.678	0.679	0.681	-0.230
IM2ELEVATION	6.501	12.403	48.534	20.245	6.762	0.759	0.764	0.772	-0.270
LUMNet	3.622	4.986	22.933	7.811	4.846	0.746	0.806	0.833	0.644
BuildMamba	1.950	2.673	13.934	4.000	2.743	0.857	0.894	0.908	0.911

The Huawei BHE dataset, characterized by complex scenes with diverse building orientations and lighting conditions, presents a rigorous testing for height prediction. As illustrated in Table 4.9, BuildMamba achieves one of the lowest overall RMSE (9.35 m) and building-only RMSE (19.50 m), outperforming DSMNet (10.41 m / 21.53 m) and

IM2HEIGHT (12.30 m / 27.99 m). Even under high-rise conditions where most models degrade, BuildMamba attains a significantly lower RMSE-HIGH of 54.67 m, compared to 74.63 m and 73.09 m for IM2HEIGHT and IM2ELEVATION, respectively. It also maintains solid performance on mid-rise (12.55 m) and low-rise (5.73 m) ranges, demonstrating resilience to urban scene complexity and imaging inconsistencies. LUMNet, in comparison, achieves an overall RMSE of 8.52 m and an R^2 of 0.385, showing balanced performance across height categories (RMSE-H: 57.18 m, RMSE-M: 11.13 m, RMSE-L: 4.13 m). Its higher δ -accuracy values ($\delta_1 = 0.818$, $\delta_2 = 0.854$, $\delta_3 = 0.887$) suggest that the integration of auxiliary building mask inputs helps maintain structural consistency, particularly for low-rise regions. However, despite its advantage in correlation and δ metrics, LUMNet’s reliance on additional semantic inputs limits its applicability in practical scenarios, whereas BuildMamba remains more practical and robust for generalized height prediction tasks.

Table 4.9: *Building Height Estimation Performance on Huawei BHE Dataset.*

Models	RMSE	RMSE-B	RMSE-H	RMSE-M	RMSE-L	δ_1	δ_2	δ_3	R^2
DSMNet	10.413	21.531	62.238	12.856	4.814	0.609	0.641	0.671	-3.152
IM2HEIGHT	12.301	27.986	74.631	20.720	8.323	0.385	0.385	0.385	-0.221
IM2ELEVATION	11.841	26.721	73.092	20.604	7.322	0.792	0.798	0.792	-0.203
LUMNet	8.521	19.494	57.183	11.132	4.132	0.818	0.854	0.887	0.385
BuildMamba	9.356	19.503	54.672	12.554	5.732	0.583	0.623	0.650	0.247

Overall, BuildMamba achieves state-of-the-art performance across both tasks together with outstanding results in each category. This indicates its generalization capability across the datasets, promising results in solving the long-tailed height distribution issue in pixel-wise height regression problems for buildings.

The visual comparisons presented in Figures 4.5, 4.6, and 4.7 offer a comprehensive evaluation of building height estimation performance across the DFC23, DFC19, and

Huawei BHE datasets, respectively. Each dataset has unique characteristics that significantly influence the performance of the models.

Figure 4.5 illustrates height estimation results on the DFC19 dataset, where the ground truth features a complex industrial building with multiple roof levels and sharp height transitions. Among the models, BuildMamba delivers the most faithful reconstruction using only RGB input, accurately capturing both global structures and local variations, while preserving clear height discontinuities. LUMNet produces smooth and coherent outputs, but relies heavily on auxiliary building masks, which guide predictions through structural priors—this leads to oversmoothing and loss of subtle details, despite favorable metrics. DSMNet, in contrast, suffers from severe noise and instability near object boundaries, resulting in distorted and inconsistent height maps.

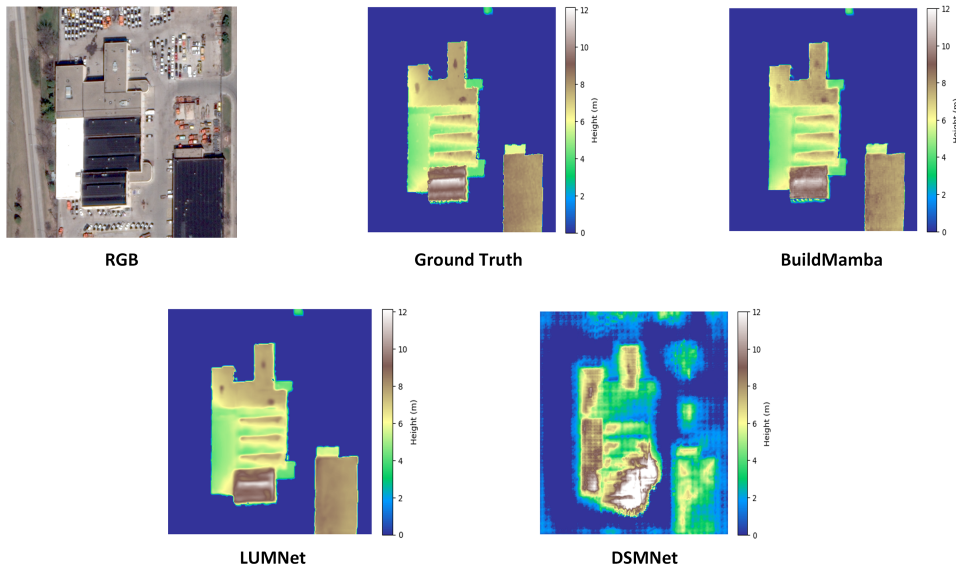


Figure 4.5: Visualization of predicted building heights on the DFC19 dataset, comparing results across competing models.

In Figure 4.6, showcasing the DFC23 dataset, we observe that BuildMamba produces height predictions that are the closest to the ground truth, especially in preserving the structure and relative height differences of high-rise buildings. This is particularly

important given that DFC23 contains 1,435 buildings over 40 meters, and a maximum height of 183 meters, which poses a challenge for conventional models. Closest competitor LUMNet introduce noisy predictions around building boundaries. Models like DSMNet tend to underestimate building heights and overly smooth details. The mean building height of 14.82 m and high standard deviation of 12.56 m in this dataset make generalization difficult, further highlighting BuildMamba’s strength in handling height variance.

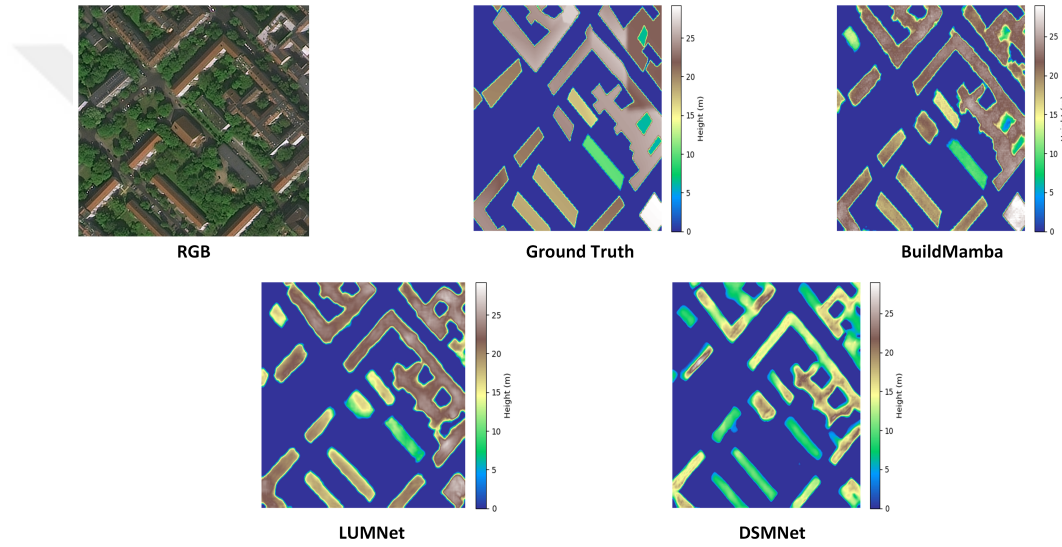


Figure 4.6: Visualization of predicted building heights on the DFC23 dataset, comparing results across competing models.

Figure 4.7, visualizing the Huawei BHE dataset, is especially significant due to its higher diversity and urban complexity. With 5,816 buildings over 40 meters, a maximum building height of 568 meters, and a very high standard deviation (21.51 m), it poses substantial generalization challenges. Even under such extreme conditions, BuildMamba and LUMNet are the models that most accurately approximate building heights while preserving structural boundaries. However, when focusing specifically on high-rise structures—as seen in both the visual results and RMSE-H values in Table 4.9—BuildMamba

is the only model capable of reliably predicting tall buildings. DSMNet either flatten the elevation range or miss finer details completely. These results reflect the effectiveness of the Sobel-based boundary-aware loss and the MAM in improving edge-level prediction and overall feature refinement.

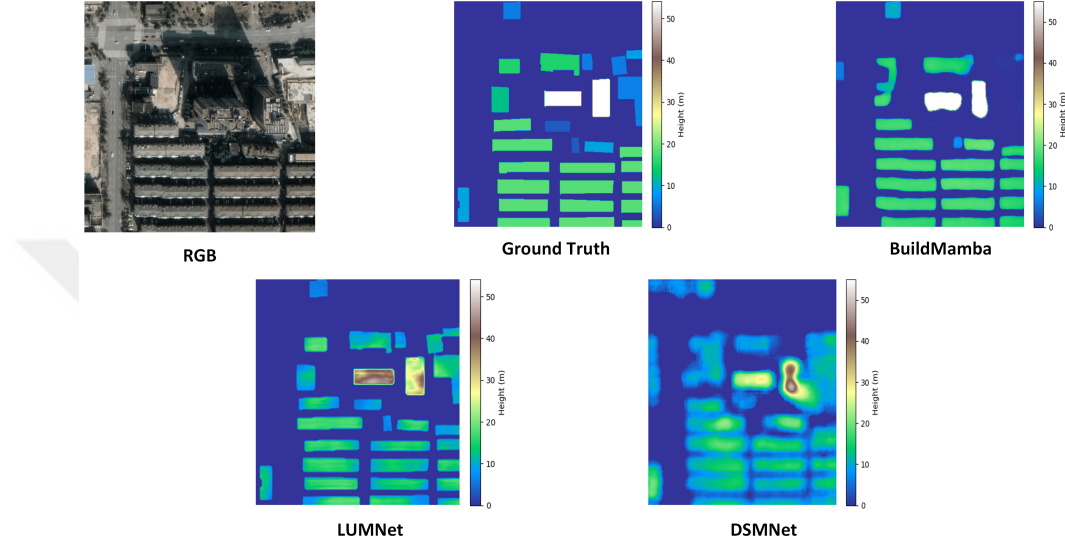


Figure 4.7: Visualization of predicted building heights on the Huawei BHE dataset, comparing results across competing models.

In conclusion, across all datasets and visualizations, BuildMamba consistently provides the most faithful reconstructions of height maps, particularly excelling in complex and high-rise scenarios. This aligns with the statistical distributions of each dataset and demonstrates the model’s superior generalization ability and adaptability to diverse urban morphologies.

4.5 Discussion and Ablation Study

To better understand the contributions of key architectural components and loss function components within our proposed framework, extensive ablation study is conducted on the DFC23 dataset. Table 4.10 lists various configurations and Table 4.11 summarizes the performance of these tested configurations, including the integration of

Mamba Attention Module (MAM), Adaptive Binning (AdaBins), and Boundary-Aware Loss with Sobel operator and weighting scheme of tall buildings, along with their combinations. The evaluation focuses on both semantic segmentation metrics (IoU, F1) and building height estimation accuracy across different building height categories.

In order to evaluate the effect of different head architectures on the performance of the proposed multi-task learning framework, a series of experiments were conducted by varying both the segmentation and height estimation heads. For the semantic segmentation task, two widely adopted architectures ASPP and UPerNet were examined. ASPP [7] leverages multiple parallel dilated convolutions with varying receptive fields to effectively capture multi-scale contextual information, which is particularly beneficial for detecting objects of different sizes in complex urban scenes. On the other hand, UPerNet integrates FPN structure that fuses hierarchical feature maps from multiple levels of the backbone, thereby enhancing spatial understanding and boundary precision.

For the height estimation task, Adaptive Binning (AdaBins) and VPDepth heads were compared. AdaBins [5] dynamically predicts bin centers for depth (or height) estimation, allowing adaptive discretization of continuous height values and improving performance on regions with diverse elevation patterns. VPDepth [54], in contrast, incorporates vision priors and lightweight decoding layers inspired by generative diffusion-based models, focusing on efficient yet accurate reconstruction of continuous height maps. By systematically analyzing the combinations of these heads, the study aims to identify the most effective configuration that balances semantic accuracy, and height precision.

4.5.1 Impact of Proposed Modules on Model Performance

Table 4.10 provides an overview of the experimental configurations used to evaluate the effectiveness of the proposed modules within the framework. Specifically, three

main components are investigated: MAM, AdaBins, and the Boundary-Aware Loss function (B.A. Loss). Each configuration is designed to isolate and assess the contribution of individual modules as well as their combinations. Model 0 serves as the baseline configuration without incorporating any additional modules. Models 1–3 examine the independent influence of each module when applied separately, while Models 4–6 focus on pairwise integrations to observe potential synergies. Finally, Model 7 incorporates all three components simultaneously to evaluate the cumulative impact of the full configuration. This structured experimental setup allows for a comprehensive understanding of how each element affects both segmentation accuracy and height estimation performance.

Table 4.10: Configuration overview of models based on MAM, Adaptive Binning, and Boundary-Aware Loss.

Model	MAM	AdaBins	B.A. Loss
Model 0			
Model 1	✓		
Model 2		✓	
Model 3			✓
Model 4	✓	✓	
Model 5	✓		✓
Model 6		✓	✓
Model 7	✓	✓	✓

Table 4.11 presents a comprehensive ablation analysis conducted on the DFC23 dataset to assess the contribution of each proposed module. The inclusion of MAM primarily targets the hierarchical context aggregation challenge by refining spatial feature representation across varying scales. As shown in Table 4.11, when MAM is applied independently (*Model 1*), the network achieves a notable improvement in semantic consistency with an IoU of 0.924 and an F1 score of 0.960, outperforming other single-module configurations.

When MAM is combined with AdaBins (*Model 4*), a slight reduction in RMSE

Table 4.11: Ablation study on the impact of MAM, Adaptive Binning, and Loss Functions on DFC23 dataset.

Model	IoU	F1	RMSE	RMSE-B	RMSE-H	RMSE-M	RMSE-L	δ_1	δ_2	δ_3	R^2
Model 0	0.905	0.940	2.150	3.141	14.710	4.440	3.190	0.802	0.846	0.881	0.841
Model 1	0.924	0.960	1.980	2.710	14.010	4.020	2.750	0.855	0.890	0.905	0.910
Model 2	0.902	0.948	2.470	3.460	15.820	5.130	3.890	0.708	0.740	0.770	0.730
Model 3	0.922	0.959	1.960	2.690	13.940	4.010	2.750	0.857	0.891	0.904	0.910
Model 4	0.910	0.950	2.300	3.210	15.950	4.760	3.350	0.740	0.780	0.820	0.780
Model 5	0.927	0.962	1.950	2.673	13.934	4.000	2.743	0.857	0.894	0.908	0.911
Model 6	0.927	0.962	1.970	2.690	13.960	4.000	2.740	0.856	0.890	0.899	0.909
Model 7	0.911	0.953	2.280	3.180	15.860	4.750	3.300	0.770	0.798	0.830	0.810

(2.30 m) and marginal improvements in segmentation metrics (IoU: 0.910, F1: 0.950) are observed compared to the variant where only AdaBins is utilized. These results suggest that the contextual attention provided by MAM enhances AdaBins’ ability to discretize elevation gradients more effectively, particularly for mid- and low-rise buildings where fine-grained contextual understanding is crucial.

The best-performing configuration (*Model 5*) integrates MAM with the Boundary-Aware Loss, yielding the lowest RMSE (1.95 m) and the highest segmentation metrics (IoU: 0.927, F1: 0.962). This highlights a strong synergy between MAM’s multi-scale spatial reasoning and the loss function’s ability to preserve structural boundaries. In contrast, incorporating all three components simultaneously (*Model 7*) did not lead to additional improvements beyond the MAM + Boundary-Aware Loss setup. The results for *Model 7* (RMSE: 2.28 m, RMSE-B: 3.18 m, IoU: 0.911) in Table 4.11 suggest potential redundancy or interference between AdaBins’ bin-based discretization and the boundary-focused optimization, possibly due to quantization noise. Overall, the findings indicate that combining MAM with precision-oriented loss functions is more effective than stacking all modules simultaneously, offering the best trade-off between semantic accuracy and height estimation precision.

4.5.2 Loss Function Variants

The results presented in Table 4.12 focus on the effectiveness of different height loss functions in predicting building heights on the DFC23 dataset. Among the tested configurations, L1 Loss achieves the lowest overall RMSE (1.94), indicating strong average performance across all buildings. However, the Boundary-Aware (B.A.) Loss delivers a more balanced result across height ranges, with competitive RMSE values in low (2.74), mid (4.00), and high (13.93) building categories. In contrast, Huber Loss consistently performs worse, especially in predicting high-rise buildings, where it records the highest RMSE-H value (14.20). Although L1 Loss achieves marginally lower error, the B.A Loss also delivers stable performance across all height intervals, underscoring its suitability for real-world building height estimation. Moreover, as seen from the visual results in the previous section, its emphasis on structural discontinuities produces more realistic nDSM predictions with noticeably sharper and more coherent building boundaries.

Table 4.12: Ablation study on the impact of Different Height Loss Functions on DFC23 dataset.

Configurations	IoU	F1	RMSE	RMSE-B	RMSE-H	RMSE-M	RMSE-L	δ_1	δ_2	δ_3	R^2
Huber Loss	0.922	0.948	2.023	2.814	14.201	4.123	2.867	0.858	0.892	0.906	0.908
L1 Loss	0.924	0.961	1.944	2.653	13.901	3.965	2.708	0.859	0.900	0.921	0.913
B.A Loss	0.927	0.962	1.950	2.673	13.934	4.000	2.743	0.857	0.894	0.908	0.911

4.5.3 Comparison of Backbones

In this section, three different backbone architectures—ViT which is the backbone of the BuildFormer [3], Swin-S, and Mamba-S—are compared to evaluate their feature extraction capabilities within the proposed multi-task learning framework on the DFC23 dataset. The primary objective of this comparison is to analyze how different architectural paradigms, namely the transformer-based global attention (ViT), hierarchical

shifted-window design (Swin-S), and state-space sequence modeling (Mamba-S), influence both segmentation and height estimation performance.

As summarized in Table 4.13, the proposed Mamba-S backbone achieves the best results across all evaluation metrics. It obtains the highest IoU of 0.927 and F1-score of 0.962, outperforming both ViT (IoU: 0.843, F1: 0.902) and Swin-S (IoU: 0.861, F1: 0.925). In terms of height prediction accuracy, Mamba-S yields the lowest RMSE (1.95) and RMSE-B (2.67), indicating superior precision in overall and building-region height estimations. Moreover, it consistently reduces the errors across all height intervals—RMSE-L: 2.74, RMSE-M: 4.00, and RMSE-H: 13.93—demonstrating better adaptability to varying building scales from low-rise to high-rise structures.

The δ accuracy metrics further support these findings, where Mamba-S achieves $\delta_1 = 0.857$, $\delta_2 = 0.894$, and $\delta_3 = 0.908$, surpassing both Swin-S and ViT in prediction reliability. Additionally, the R^2 value of 0.91 indicates a strong correlation between predicted and ground truth heights, significantly higher than ViT (0.798) and Swin-S (0.815). These improvements reflect the Mamba-S backbone’s ability to effectively capture both spatial and contextual dependencies, leveraging its selective state-space representation to model long-range dependencies with high computational efficiency.

Overall, the results demonstrate that Mamba-S not only enhances the representational power for fine-grained segmentation but also provides superior generalization in height estimation, making it a highly efficient and accurate backbone for multi-task learning in remote sensing applications.

Table 4.13: *Comparison of different Backbone architectures on the DFC23 dataset.*

Backbones	IoU	F1	RMSE	RMSE-B	RMSE-H	RMSE-M	RMSE-L	δ_1	δ_2	δ_3	R^2
ViT	0.843	0.902	2.965	4.417	24.469	7.010	3.952	0.745	0.785	0.80	0.798
Swin-S	0.861	0.925	2.872	4.204	21.226	6.578	3.511	0.694	0.757	0.782	0.815
Mamba-S	0.927	0.962	1.950	2.673	13.934	4.000	2.743	0.857	0.894	0.908	0.911

4.5.4 Comparison of Different Segmentation Heads

In this section, three different segmentation heads are compared to evaluate their effectiveness on the DFC23 dataset: ASPP, UPerNet, and the proposed head (Ours). The purpose of this comparison is to analyze how different decoder structures influence both the semantic segmentation quality and the height estimation performance within the multi-task learning framework. Each head was trained under the same experimental settings using identical backbone and loss configurations to ensure a fair evaluation.

As presented in Table 4.14, the proposed segmentation head achieves the best performance across all metrics. Specifically, it attains the highest IoU of 0.927 and F1-score of 0.962, outperforming ASPP (IoU: 0.840, F1: 0.911) and UPerNet (IoU: 0.861, F1: 0.925). In terms of height estimation accuracy, the proposed head also demonstrates the lowest RMSE (1.95) and RMSE-B (2.67), indicating superior consistency between predicted and ground truth building heights. Furthermore, it maintains lower errors across all height intervals (RMSE-L: 2.74, RMSE-M: 4.00, RMSE-H: 13.93) compared to the other two configurations, highlighting its balanced capability to generalize over low-, mid-, and high-rise structures.

Using different segmentation heads leads to substantial differences in height estimation performance because the segmentation output directly shapes the spatial priors used by the height head. A stronger segmentation head provides cleaner building boundaries, more accurate footprint masks, and sharper structural cues, which guide the height predictor to better localize building regions and reduce boundary ambiguity. In contrast, weaker segmentation heads introduce noisy or imprecise masks, causing the height branch to misinterpret building extents—particularly for tall structures—thereby increasing RMSE and degrading δ accuracy. As a result, improvements in the segmentation head translate into more reliable geometric structure cues, ultimately yielding significantly better height predictions.

Regarding the δ accuracy metrics, the proposed segmentation head achieves the best results with $\delta_1 = 0.857$, $\delta_2 = 0.894$, and $\delta_3 = 0.908$, signifying a higher percentage of predictions falling within acceptable relative error thresholds. Additionally, the coefficient of determination ($R^2 = 0.91$) further confirms the strong correlation between predicted and actual heights, substantially outperforming UPerNet (0.82) and ASPP (0.89).

Overall, the results demonstrate that the proposed head provides a more effective feature aggregation and multi-scale representation than ASPP or UPerNet, allowing it to better capture complex urban details while maintaining robust height prediction accuracy.

Table 4.14: *Comparison of Different Segmentation Heads on the DFC23 Dataset.*

Seg. Head	IoU	F1	RMSE	RMSE-B	RMSE-H	RMSE-M	RMSE-L	δ_1	δ_2	δ_3	R^2
ASPP	0.840	0.911	2.131	2.963	15.076	4.628	2.922	0.831	0.876	0.892	0.890
UPerNet	0.861	0.925	2.873	4.202	21.226	6.573	3.519	0.694	0.757	0.782	0.815
Ours	0.927	0.962	1.950	2.673	13.934	4.000	2.743	0.857	0.894	0.908	0.911

4.5.5 Comparison of Height Estimation Heads

Table 4.15 presents a comparative analysis of different height estimation heads evaluated on the DFC23 dataset. The objective of this experiment is to assess how varying decoding strategies influence the model’s ability to recover accurate and structurally consistent building heights while maintaining semantic coherence in segmentation.

The first approach using *AdaBins* exhibits limited performance, achieving an RMSE of 2.47 m and an IoU of 0.902. Although *AdaBins* adaptively discretizes continuous height values into learned bin centers, its reliance on global bin prediction can lead to quantization artifacts and difficulties in modeling fine-grained variations, particularly in high-rise regions (RMSE-H: 15.82 m). This indicates that while *AdaBins* effectively captures general elevation trends, it struggles to preserve localized structural precision.

The *VPDDepth* head extends the idea of depth estimation by incorporating diffusion-inspired priors derived from vision-language generative models. As shown in Table 4.15,

it achieves slightly higher segmentation performance than AdaBins (IoU: 0.905, F1: 0.950) but underperforms in height regression metrics (RMSE: 2.53 m, RMSE-H: 17.11 m). The model benefits from global context propagation through learned priors, which improves the overall spatial smoothness of predictions.

Table 4.15: *Comparison of Different Height Heads on the DFC23 dataset.*

Height Head	IoU	F1	RMSE	RMSE-B	RMSE-H	RMSE-M	RMSE-L	δ_1	δ_2	δ_3	R^2
Adabins	0.902	0.948	2.472	3.464	15.823	5.131	3.898	0.708	0.742	0.771	0.732
VPDDepth	0.905	0.950	2.534	3.601	17.113	5.314	4.001	0.681	0.714	0.762	0.717
Ours	0.927	0.962	1.950	2.673	13.934	4.000	2.743	0.857	0.894	0.908	0.911

In contrast, the proposed height estimation head (*Ours*) demonstrates substantial improvement across all metrics (IoU: 0.927, F1: 0.962, RMSE: 1.95 m, R^2 : 0.91), significantly outperforming both baselines. As seen in Table 4.15, it achieves the highest IoU (0.927), F1 score (0.962), and a significantly lower RMSE (1.95 m). Furthermore, the δ_1 , δ_2 , and δ_3 accuracy values (0.857, 0.894, and 0.908, respectively) and the high coefficient of determination ($R^2 = 0.91$) confirm the robustness and precision of the predicted height maps, and highlight the prediction stability across diverse building heights. The proposed model effectively captures both local details and global elevation context.

5. CONCLUSION AND FUTURE WORK

In this thesis, BuildMamba, a dual-path encoder-based multi-task learning network, is proposed to address the key challenges associated with building segmentation and height estimation from monocular satellite imagery. Prior studies frequently struggle with blurred or biased height predictions due to the dominance of background pixels, the underestimation of tall buildings in complex urban environments, and the inherent difficulty of inferring accurate height information from RGB images alone, which lack explicit geometric cues. BuildMamba is designed to overcome these limitations through a joint learning architecture that couples building segmentation with height estimation, enabling the two tasks to reinforce one another. To improve the reconstruction of height discontinuities and structural boundaries, the framework incorporates a boundary-aware L1 loss, which increases the contribution of pixels along building edges and those belonging to structures taller than 40 meters. Additionally, the introduction of the Mamba Attention Module (MAM) enhances spatial feature refinement by strengthening the global contextual information captured by the VMamba backbone.

Extensive experiments on the DFC19, DFC23, and Huawei BHE datasets demonstrate that BuildMamba achieves state-of-the-art performance, consistently outperforming competitive baselines in both segmentation accuracy and height estimation. These results validate the strong generalization capability of the proposed model across diverse urban morphologies and highlight the effectiveness of its architectural design and loss formulation in addressing challenges in monocular height prediction.

As future work, several promising directions can further enhance the performance and scalability of the proposed framework. One avenue is to increase the representation of tall buildings within the training dataset, as their scarcity contributes to the difficulty of learning accurate height distributions. This imbalance can be addressed by employing diffusion models to synthesize realistic image-mask-height map triplets, thus creating

a more balanced and diverse dataset. Such generative augmentation could improve the model’s robustness and reduce bias toward low-rise and mid-rise structures.

Additionally, integrating auxiliary modalities offers substantial potential for boosting height estimation accuracy. For example, incorporating shadow information, sun-angle metadata, or SAR imagery may help resolve ambiguities in monocular predictions, particularly for skyscrapers or densely occluded areas. Encoding these modalities either through parallel branches or attention-based fusion could provide richer geometric cues and reduce reliance on RGB-only textures. Beyond these additions, exploring temporal data, multi-view consistency, or self-supervised pretraining strategies may further increase generalization across diverse cities and imaging conditions. Collectively, these extensions represent meaningful steps toward developing more reliable, scalable, and multimodal building height estimation systems suitable for real-world remote sensing applications.

REFERENCES

- [1] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, “Vmamba: Visual state space model,” *Advances in neural information processing systems*, vol. 37, pp. 103031–103063, 2024.
- [2] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.
- [3] L. Wang, S. Fang, X. Meng, and R. Li, “Building extraction with vision transformer,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.
- [4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.
- [5] S. F. Bhat, I. Alhashim, and P. Wonka, “Adabins: Depth estimation using adaptive bins,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4009–4018, 2021.
- [6] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, “Unified perceptual parsing for scene understanding,” in *Proceedings of the European conference on computer vision (ECCV)*, pp. 418–434, 2018.
- [7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
- [8] J. Lao, C. Wang, X. Zhu, X. Xi, S. Nie, J. Wang, F. Cheng, and G. Zhou, “Retrieving building height in urban areas using icesat-2 photon-counting lidar data,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 104, p. 102596, 2021. Art. no. 102596.
- [9] X. Ma, G. Zheng, X. Chi, L. Yang, Q. Geng, J. Li, and Y. Qiao, “Mapping fine-scale building heights in urban agglomeration with spaceborne lidar,” *Remote Sensing of Environment*, vol. 285, p. 113392, 2023. Art. no. 113392.
- [10] Y. Wang, Z. Xiong, C. Liu, A. J. Stewart, T. Dujardin, N. I. Bountos, A. Zavras, F. Gerken, I. Papoutsis, L. Leal-Taixé, and X. X. Zhu, “Towards a unified copernicus foundation model for earth vision,” 2025.

- [11] S. Chen, Y. Shi, Z. Xiong, and X. X. Zhu, "Htc-dc net: Monocular height estimation from single remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–18, 2023.
- [12] Y. Wang, H. Tang, Y. Zhou, and G. Wang, "Henet: A height estimation network for estimating the height of buildings from single optical image," in *2024 International Conference on New Trends in Computational Intelligence (NTCI)*, pp. 307–311, IEEE, 2024.
- [13] S. Du, J. Xing, S. Wang, X. Xiao, J. Li, and H. Liu, "Lumnet: Land use knowledge guided multiscale network for height estimation from single remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.
- [14] J. B. Campbell and R. H. Wynne, *Introduction to remote sensing*. Guilford press, 2011.
- [15] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Systems with Applications*, vol. 169, p. 114417, 2021.
- [16] L. L. Byrnside III, "Shadow-based automatic building height estimation from high spatial resolution satellite imagery," 2022.
- [17] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE transactions on knowledge and data engineering*, vol. 34, no. 12, pp. 5586–5609, 2021.
- [18] W. Xu, Z. Feng, Q. Wan, Y. Xie, D. Feng, J. Zhu, and Y. Liu, "Building height extraction from high-resolution single-view remote sensing images using shadow and side information," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.
- [19] L. Mou and X. X. Zhu, "Im2height: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," *arXiv preprint arXiv:1802.10249*, 2018.
- [20] C.-J. Liu, V. A. Krylov, P. Kane, G. Kavanagh, and R. Dahyot, "Im2elevation: Building height estimation from single-view aerial imagery," *remote sensing*, vol. 12, no. 17, p. 2719, 2020.
- [21] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2002–2011, 2018.
- [22] Q. Li, L. Mou, Y. Hua, Y. Shi, S. Chen, Y. Sun, and X. X. Zhu, "3dcentripetalnet: Building height retrieval from monocular remote sensing imagery," *International Journal of Applied Earth Observation and Geoinformation*, vol. 120, p. 103311, 2023.

- [23] Z. Li, X. Wang, X. Liu, and J. Jiang, “Binsformer: Revisiting adaptive bins for monocular depth estimation,” *IEEE Transactions on Image Processing*, 2024.
- [24] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, “Segment anything,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.
- [25] J. Wu, M. Yuan, T. Wang, X. Jia, and D.-M. Yan, “Heightformer: Single imagery height estimation transformer with bilateral feature pyramid fusion,” *IEEE Geoscience and Remote Sensing Letters*, 2024.
- [26] Q. Gao, M. Liu, C. Wang, and T. Wang, “Attention based multi-scale network for height estimation,” in *2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE)*, pp. 1211–1215, IEEE, 2025.
- [27] W. Shi, Q. Meng, L. Zhang, J. Wang, T. Zhou, and P. M. Atkinson, “Building height extraction from monocular off-nadir satellite sensor imagery,” in *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2537–2540, IEEE, 2024.
- [28] W. Zhao, C. Persello, and A. Stein, “Semantic-aware unsupervised domain adaptation for height estimation from single-view aerial images,” *ISPRS journal of photogrammetry and remote sensing*, vol. 196, pp. 372–385, 2023.
- [29] Z. Xiong, S. Chen, Y. Shi, and X. X. Zhu, “Self-supervised pre-training with monocular height estimation for semantic segmentation,” *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
- [30] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “Segformer: Simple and efficient design for semantic segmentation with transformers,” in *Neural Information Processing Systems (NeurIPS)*, 2021.
- [31] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, “Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4096–4105, 2020.
- [32] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, “Farseg++: Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13715–13729, 2023.
- [33] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, “Swin transformer embedding unet for remote sensing image semantic segmentation,” *IEEE transactions on geoscience and remote sensing*, vol. 60, pp. 1–15, 2022.

- [34] X. Li, H. He, X. Li, D. Li, G. Cheng, J. Shi, L. Weng, Y. Tong, and Z. Lin, "Point-flow: Flowing semantics through points for aerial image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4217–4226, 2021.
- [35] J. Li, W. He, W. Cao, L. Zhang, and H. Zhang, "Uanet: An uncertainty-aware network for building extraction from remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.
- [36] H. Guo, B. Du, L. Zhang, and X. Su, "A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 183, pp. 240–252, 2022.
- [37] P. Zhu, Z. Song, J. Liu, J. Yan, X. Luo, and Y. Tao, "Mshformer: a multiscale hybrid transformer network with boundary enhancement for vhr remote sensing image building extraction," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.
- [38] L. Xu, Y. Li, J. Xu, Y. Zhang, and L. Guo, "Bctnet: Bi-branch cross-fusion transformer for building footprint extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.
- [39] H. Guo, X. Su, C. Wu, B. Du, and L. Zhang, "Decoupling semantic and edge representations for building footprint extraction from remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.
- [40] H. Zhang, Y. Liao, H. Yang, G. Yang, and L. Zhang, "A local–global dual-stream network for building extraction from very-high-resolution remote sensing images," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 3, pp. 1269–1283, 2020.
- [41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [42] X. Ma, X. Zhang, and M.-O. Pun, "Rs 3 mamba: Visual state space model for remote sensing image semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.
- [43] Z. Zheng, Y. Zhong, and J. Wang, "Pop-net: Encoder-dual decoder for semantic segmentation and single-view height estimation," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4963–4966, IEEE, 2019.
- [44] M. Elhousni, Z. Zhang, and X. Huang, "Height prediction and refinement from aerial images with semantic and geometric guidance," *IEEE Access*, vol. 9, pp. 145638–145647, 2021.

- [45] J. Mahmud, T. Price, A. Bapat, and J.-M. Frahm, “Boundary-aware 3d building reconstruction from a single overhead image,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 441–451, 2020.
- [46] Y. Mao, X. Sun, X. Huang, and K. Chen, “Light: Joint individual building extraction and height estimation from satellite images through a unified multitask learning network,” in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5320–5323, IEEE, 2023.
- [47] Y. Chen, Q. Yan, and W. Huang, “Mftsc: A semantically constrained method for urban building height estimation using multiple source images,” *Remote Sensing*, vol. 15, no. 23, p. 5552, 2023.
- [48] C. Lu, N. Cao, P. Zhang, T. Liu, B. Peng, G. Liu, M. Yuan, S. Zhang, S. Huang, and T. Wang, “Hgdnet: A height-hierarchy guided dual-decoder network for single view building extraction and height estimation,” in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 758–761, IEEE, 2023.
- [49] Z. Gao, W. Sun, Y. Lu, Y. Zhang, W. Song, Y. Zhang, and R. Zhai, “Joint learning of semantic segmentation and height estimation for remote sensing image leveraging contrastive learning,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.
- [50] T. Wu, R. Zhao, M. Lv, Z. Jia, L. Li, M. Liu, X. Zhao, H. Ma, and G. Vivone, “Efficient mamba-attention network for remote sensing image super-resolution,” *IEEE Transactions on Geoscience and Remote Sensing*, 2025.
- [51] X. Li, F. Xu, F. Liu, X. Lyu, Y. Tong, Z. Xu, and J. Zhou, “A synergistical attention model for semantic segmentation of remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.
- [52] D. Vidaurre, C. Bielza, and P. Larranaga, “A survey of l1 regression,” *International Statistical Review*, vol. 81, no. 3, pp. 361–387, 2013.
- [53] A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Salt Lake City, UT, USA), pp. 7482–7491, 2018.
- [54] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, “Unleashing text-to-image diffusion models for visual perception,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5729–5739, 2023.
- [55] P. J. Huber, “Robust estimation of a location parameter,” *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964. Art. no. 1177703732.

- [56] C. Persello, R. Hänsch, G. Vivone, and et al., “2023 ieee grss data fusion contest: Large-scale fine-grained building classification for semantic urban reconstruction [technical committees],” *IEEE Geoscience and Remote Sensing Magazine*, vol. 11, pp. 94–97, March 2023.

