remote sensing

MDPI

Article

Fusion of RGB and LiDAR Modalities for Building Footprint Extraction Using High-Resolution Aerial Imagery

Norbert Serbán [1,*], Péter Enyedi [2], Péter Burai [3] and Balázs Harangi [1]

[1] Faculty of Informatics, University of Debrecen, 4028 Debrecen, Hungary; harangi.balazs@inf.unideb.hu
[2] Envirosense Hungary Ltd., 4032 Debrecen, Hungary; peter.enyedi@envirosense.hu
[3] Remote Sensing Centre, University of Debrecen, 4028 Debrecen, Hungary; burai.peter@unideb.hu
* Correspondence: serban.norbert@inf.unideb.hu

Highlights

What are the main findings?

- The proposed RGB–LiDAR fusion model, Point U-Net, significantly outperforms traditional single-source (RGB-only or LiDAR-only) and ensemble segmentation methods in building detection accuracy.
- Integrating 2D image features with 3D point cloud information at multiple decoding levels leads to more detailed and reliable semantic segmentation results for urban and environmental mapping tasks.

What are the implications of the main findings?

- The improved accuracy of RGB–LiDAR fusion models suggests that combining complementary data sources can greatly enhance the reliability and accuracy of building detection, benefiting applications such as urban planning, infrastructure monitoring, and environmental analysis.
- This approach demonstrates the potential for developing more advanced multimodal deep learning architectures, encouraging further research into data fusion techniques for other remote sensing and geospatial analysis tasks.

Abstract

In this paper, a novel approach is presented for fusing RGB and LiDAR inputs for semantic segmentation. Accurate building detection is required for various scenarios such as urban planning or environmental monitoring. The two main sources for accurate building segmentation are either RGB aerial images or LiDAR point clouds covering the selected area. Each of these sources has its own well-known techniques for segmentation; however, for the combination of the input, there are not many architectures available, and extracting different features from the two different fields can result in an enhanced segmentation map. The authors of this article created a semantic segmentation model that uses both the aerial RGB image and the LiDAR point cloud as its input. The network first takes the point cloud and forwards the processed projection to a modified U-Net-based architecture, which fuses the extracted features of the 3D input with the extracted information of the 2D input on each level of the decoding. To train and test the presented model, the authors used a dataset containing more than 3000 images and their corresponding 3D point clouds of three different areas from Hungary. As is also presented in this paper, this approach provides significantly better results than the traditional RGB, Point Cloud segmentation models, and their ensembles in terms of segmentation accuracy.

Academic Editor: Qiangqiang Yuan

Received: 1 April 2026
Revised: 31 May 2026
Accepted: 17 June 2026
Published: 21 June 2026

Keywords: semantic segmentation; aerial imagery; multimodal segmentation; high resolution imagery

1. Introduction

Building detection is a fundamental task in urban planning, disaster management, and environmental monitoring, with significant applications in areas such as land-use mapping, infrastructure assessment, and urban expansion analysis [1,2]. Over the years, advancements in remote sensing technologies, particularly the utilization of LiDAR (Light Detection and Ranging) data, have revolutionized this domain. LiDAR, first introduced in the 1960s for topographic mapping, has evolved significantly, enabling highly accurate 3D spatial information acquisition and structural analysis [3]. Its ability to penetrate vegetation and provide detailed elevation models has made it indispensable for applications such as forest management, flood risk assessment, and urban landscape characterization [4].

When combined with RGB imagery, which captures high-resolution spectral information, the fusion of these two data sources enables robust and precise building detection [5]. RGB data provides detailed texture and color information, while LiDAR adds critical depth and structural context, improving the accuracy and reliability of detection models [6]. For example, studies have successfully employed this fusion for automated building extraction in urban environments [7,8] and damage assessment in post-disaster scenarios [9,10]. In the next sections, we briefly describe and summarize the past and present of the task of semantic segmentation by looking at the unimodal approaches and the solutions based on multimodality.

1.1. RGB Image Segmentation

Since the original U-Net architecture was introduced in 2015 [11], it has become immensely popular among research groups as its lightweight architecture and effortless customization of encoders facilitate many applications of the model in terms of semantic segmentation on a wide variety of topics, as well as aerial imagery. The article by [12] introduced an enhanced U-Net architecture that integrates a self-attention mechanism and separable convolutions. This modification demonstrated significant improvements in urban landscape segmentation, highlighting its efficacy even years after the initial development of the original U-Net design. Their approach effectively leverages the strengths of self-attention for capturing long-range dependencies and separable convolutions for computational efficiency, making it particularly well-suited for processing high-resolution aerial imagery. In their study, the authors of ref. [13] proposed a Bayesian U-Net framework incorporating Monte Carlo dropout layers to quantify uncertainty in the semantic segmentation of Earth observation imagery. This approach not only enhances the interpretability of the segmentation outcomes by providing uncertainty estimates but also proves valuable for decision-making in applications where reliability is critical. As demonstrated by this work and others, the U-Net architecture remains a robust and versatile tool for a wide range of use cases. Despite the introduction of numerous alternative architectures and advancements in semantic segmentation, the U-Net continues to maintain its relevance, demonstrating adaptability and effectiveness across diverse domains.

The DeepLabv3+ model [14] is also a widely used semantic segmentation network, as it can perform well across a wide variety of scenarios. It merges the benefits of the encoder–decoder architecture with multi-scale feature extraction using the atrous convolution of DeepLabv3 [15]. The encoder uses atrous spatial pyramid pooling (ASPP) for high-level feature extraction to different object scales, and the decoder combines low-level features to

enhance spatial details. Its efficiency stems from the tradeoff of computational expense and segmentation precision, which is advantageous for real-time uses. DeepLabv3+ has been shown to achieve great results in image segmentation of objects, even with intricate borders. The Aerial LaneNet [16] also uses Deeplabv3+ as a basis semantic segmentation model for the task of lane segmentation from aerial images enhanced by wavelet decomposition to capture multi-scale features and improve spatial resolution. The authors of [17] present an adaptive DeepLabv3+ model for semantic segmentation of aerial images, optimized using an Improved Golden Eagle Optimization Algorithm (IGEOA). Their proposed method enhances DeepLabv3+ by tuning hyperparameters such as learning rate and batch size with IGEOA, improving segmentation performance. IGEOA incorporates adaptive strategies to balance exploration and exploitation, ensuring efficient convergence to optimal solutions.

The Pyramid Scene Parsing Network (PSPNet) [18] introduced a novel approach to semantic segmentation in 2017 by incorporating so-called pyramid pooling modules to capture global context information. By pooling features at multiple scales and combining them, PSPNet effectively integrates local and global cues to improve segmentation accuracy. The model used a fully convolutional architecture and achieved state-of-the-art performance on challenging datasets like ADE20K and Cityscapes. During that time, the model significantly advanced the field of scene understanding, making it applicable to tasks such as autonomous driving and image parsing. The study of [19] applied on UAV RGB images and an improved PSPNet to identify wheat lodging areas. The researchers applied a model based on PSPNet using MobileNetV2 as an encoder and processed the feature maps by NAM to reduce the calculation complexity of the network.

As a new approach for semantic segmentation, transformer-based architectures are introduced. The transformers, which are mainly built on the self-attention mechanism, were first applied in the field of natural language processing [20]. However, using the strong representation capabilities of the architecture, the researchers quickly realized its potential for computer vision tasks.

For the field of remote sensing and aerial image segmentation, transformer-based approaches were also published. Ref. [21] developed an architecture called AerialFormer, which uses the combination of the benefits of CNN and Transformer architectures, mainly using the Transformer as the encoder and a multi-dilated CNN as a decoder. The authors of [22] proposed a novel architecture called the crisscross-global vision transformer applied to the job of semantic segmentation of very high-resolution aerial imagery. The 2dsegformer [23], also for the same semantic segmentation task, provided 2D positional attention in their model to accurately record the 2D information from the aerial images and pass that information to the transformer model.

1.2. Point Cloud Segmentation

Apart from the RGB images, semantic segmentation is also possible on point clouds. As the point cloud represents geometric data structures, it contains a significant amount of geometric information and can illustrate complex 3D environments accurately. However, the challenges of point cloud segmentation are the lack of a fixed structure of the data, as it usually has characteristics like sparsity or randomness. As the deep learning approaches evolved over the years, many researchers became interested in point cloud segmentation as it has many applications prospects [24].

As a breakthrough in semantic segmentation of point clouds, PointNet [25] was released in 2017. This architecture directly takes point clouds as input and returns the labels for every point of the point cloud. The original setting only requires the coordinates of the points (x, y, z); however, the dimensions can be extended by additional features. By using local and global information aggregation of the input point clouds, PointNet provides

accurate prediction of per-point quantities, which enables the architecture to outperform the state-of-the-art networks on shape part segmentation and scene segmentation.

As an improvement of the original architecture, PointNet++ [26] was published in the same year. The main progress of the network was achieved in capturing the local structures of the point clouds by applying PointNet recursively on nested partitions of the original point set; therefore, the network is able to provide more efficient learning of local features with increasing contextual scales. As PointNet++ provides an excellent base for aerial scene segmentation, many applications use the architecture as the fundamental foundation of their solutions. Ref. [27] provided a solution for the segmentation of transmission corridors based on point set data given by unmanned aerial vehicles (UAVs). Their proposed CA-PointNet++ network applied the so-called Coordinate Attention mechanism on the original PointNet++ architecture, which can embed positional information into channel attention, enhancing the locations of the region of interest more accurately. The authors of [28] developed an attention mechanism system that incorporates into the sampling operations of the PointNet++ set abstraction layers. Their approach was used for building extraction on point cloud datasets captured by UAVs.

To solve the problem of point set semantic segmentation, other approaches were also implemented. The RandLa-Net [29], published in 2020, described as a lightweight neural network, used random point sampling instead of point selection approaches. To preserve geometric details, the architecture introduced a novel local feature aggregation module. By developing the random sampling method, the network achieved a significant increase in point processing capabilities compared to the existing solutions. The network also contributed to many point cloud segmentation solutions based on aerial point sets [30–32]. The KPConv [33] from 2019 operates on point sets without using any intermediate representation, as the convolutional weights of the network are located in the Euclidean space by kernel points and applied to the input points close to them. As these locations are continuous in space, they can be learned by the network and can be extended to deformable convolutions which can adapt kernel points to local geometry.

1.3. Multi-Modal Semantic Segmentation

As the task of semantic segmentation is possible on both RGB images and point clouds, it is a straightforward next step to combine the two separate modalities and use their synergy to enhance the effectiveness of the segmentation problem-solving. Recent deep learning-based approaches have explored increasingly sophisticated strategies for fusing RGB imagery and LiDAR data in semantic segmentation tasks. PMNet [34] utilized a detailed 3D semantic segmentation of urban scenes using the fusion of point-wise LiDAR point cloud segmentation and image segmentation. As per the architecture of the network, the authors combined a PointNet and a CNN encoder–decoder. To fuse the features of the modalities, the authors created a spatial correspondence table to match the coordinates of the point set with the pixels of the input image.

The other approach to combine the RGB and LiDAR inputs is made by the authors of Direct LiDAR-Aerial Fusion Network (DLAFNet) [35]. The network was built on KPConv network for the point cloud segmentation and uses Transformer Blocks to process the RGB image input. The network feeds the result of each KPConv step into the features created by the Transformer Block by projecting the point set features onto the RGB features using the base coordinates. The fused layers are then processed by the MLP module and concatenated after each step to receive the predicted segmentation map of the inputs. A simplified way to handle LiDAR point clouds is to project them to the 2D space and create a depth feature map of the 3D point cloud. The EDFT [36] uses this way to preprocess 3D point sets into a simplified format which reduces the memory usage and computation

costs. The researchers implemented a depth-aware self-attention (DSA) module to mitigate the gap between the depth image and the RGB input by fusing the depth feature with the color features. The DSA module is applied to the encoder phase of the network, and the decoding is done on the merged feature maps on every upsampling layer. The mapping of the high-dimensional features to low-dimensional features during or before the feature extraction can inevitably lead to information loss. The Imbalance Knowledge-Driven Multi-modal Network (IKD-Net) [37] is capable of mining imbalance across the modalities and using the strong modal to drive the refinement of the weaker feature map. To achieve this modality-aware refinement, the authors introduced two dedicated components: the Global Knowledge-Guided module and the Class Knowledge-Guided module. The first module operates at a global scale, capturing cross-modal dependencies and transferring broad contextual cues from the dominant modality to enhance the other. In contrast, the Class Knowledge-Guided module focuses on category-specific cues, ensuring that fine-grained class-level information is effectively shared across modalities. Together, these modules form a complementary mechanism that enhances both global representation quality and class-level discriminability. These modules are then integrated into architecture which is based on U-Net and RandLA-Net networks.

In this paper, we would like to propose a fusion of point cloud and image semantic segmentation networks which take the advantages of both architectures by taking spatial knowledge from the point clouds and enhancing it with the textual-based information extracted from the image semantic segmentation network. By that, we would expect to significantly increase the segmentation accuracy of urban regions on a multimodal aerial dataset. The above-mentioned methods demonstrate that carefully designed fusion strategies can outperform simple early- or late-fusion baselines and highlight the importance of multimodal feature interaction for accurate semantic segmentation. Despite their effectiveness, existing multimodal architectures such as PMNet and DLAFNet primarily operate on image-aligned feature grids and rely on fixed fusion schemes within the network. As a result, they do not explicitly leverage point-level sampling during feature decoding, nor do they investigate decoder architectures that preserve modality-specific representations until late fusion stages. In contrast, our proposed framework explicitly incorporates point-based feature sampling from LiDAR data and employs a multi-branch decoder design that enables complementary processing of RGB imagery, depth information, and sampled point-cloud activations. This design aims to improve robustness to point density variations and spatial misalignment while preserving fine-grained structural details of building footprints. Based on these considerations, PMNet and DLAFNet are selected as representative state-of-the-art multimodal baselines in the experimental evaluation presented in Section 3.

A notable challenge in this study arises from the use of standard orthophotos rather than true orthophotos. Standard orthophotos suffer from relief displacement, leading to perspective distortion of elevated features such as buildings. Consequently, a spatial misalignment exists between the RGB imagery and the geometrically accurate LiDAR data, despite simultaneous acquisition. This geometric mismatch introduces noise in the multimodal fusion process, which can negatively impact the accuracy of automated building extraction. It is important to address the geometric discrepancies between the two modalities used in this research. Since standard orthophotos were utilized instead of true orthophotos, the aerial imagery exhibits inherent perspective distortion, commonly known as building lean. In contrast, the LiDAR data provides highly accurate, orthographic 3D geometry. This fundamental difference results in a spatial offset between the building roofs in the RGB imagery and their corresponding locations in the LiDAR data. This co-registration error poses a significant challenge for object recognition algorithms, as the spatial inconsistencies between the fused features can degrade the boundary delineation

and overall extraction performance. This misalignment acts as conflicting information during the feature fusion stage, complicating the network's ability to learn accurate multi-modal representations for building footprint extraction. Although this misalignment can negatively affect multimodal segmentation, the proposed methodology contains a point-level feature sampling strategy which introduces spatial tolerance by aggregating local point neighborhoods, thereby reducing the sensitivity of the fusion process to small spatial offsets.

Altogether, the primary objective of this study is to improve building footprint extraction from high-resolution aerial data by explicitly combining point-level geometric information from LiDAR point clouds with texture-rich features extracted from RGB imagery. To this end, the main contributions of this work are threefold: the proposal of a novel Point U-Net architecture that integrates point-based LiDAR feature sampling and multi-branch decoder-level fusion with a U-Net-based image segmentation framework, a systematic ablation study that quantitatively analyzes the impact of LiDAR sampling strategies, fusion components, and decoder design choices, and a comprehensive experimental evaluation against unimodal baselines and recent state-of-the-art RGB–LiDAR fusion methods on a high-resolution UAV dataset.

2. Materials and Methods

2.1. Dataset

We compiled a comprehensive multi-modal dataset that integrates high-resolution RGB imagery with LiDAR point-cloud data, all collected across three distinct urban environments in Hungary. The data acquisition process was carried out using unmanned aerial vehicles (UAVs) equipped with a multisensory payload, enabling the simultaneous capture of complementary two-dimensional visual information and three-dimensional structural measurements. This coordinated collection strategy ensured that both modalities were spatially and temporally aligned, which is essential for downstream tasks involving cross-modal fusion or comparative analysis. Although standard orthophotos are used as RGB inputs to the network, these products are generated from aerial images acquired simultaneously with the LiDAR data; thus, the term "orthophoto" refers to the processing level rather than a separate data acquisition. The overall study area is divided into three separate spatial extents: two smaller sub-regions covering 0.19 km^2 and 0.68 km^2, respectively, and a considerably larger urban zone exceeding 5 km^2. These regions were selected to represent a diverse mixture of building densities, architectural patterns, and landscape characteristics, thereby enhancing the robustness and generalizability of the dataset. All annotations were produced manually by trained domain experts. During the labeling process, annotators made use of both the LiDAR point clouds and the RGB imagery to ensure high-quality, geometry-aware delineation of building footprints. The finalized annotations were stored as shapefiles, which serve as the foundation for generating supervision signals for machine-learning models. Since our research relies on a binary semantic segmentation framework, we converted the shapefiles into binary masks corresponding to the two classes defined in the dataset: class 0 representing background and class 1 representing building structures. These masks were generated for both the image tiles and the point-cloud partitions to maintain consistency between modalities.

To prepare the RGB imagery for training and evaluation, the raw images were divided into tiles of 512 $\times$ 512 pixels with a 5% overlap between adjacent tiles, which helps mitigate boundary artifacts during segmentation. All RGB frames were fully georeferenced, allowing for precise calculation of inter-pixel distances and accurate alignment with the LiDAR data. This geospatial consistency is crucial when projecting the imagery onto the point cloud or when comparing model outputs across modalities. The pixel spacing in the

imagery was 30 cm, and for this reason, the LiDAR point clouds were similarly tiled into matching 512 × 512 regions, each associated with the corresponding binary mask. Due to the relatively high density of the LiDAR data used in this research, we further processed the imagery to ensure adequate point-to-pixel correspondence. Specifically, the image tiles were downscaled to 128 × 128 pixels so that each pixel would contain at least one LiDAR point, guaranteeing meaningful cross-modal associations for training. This adjustment improved the reliability of the dataset when used for multimodal learning tasks and ensured that each pixel-level label could be supported by corresponding three-dimensional information.

Ultimately, the dataset comprises more than 3300 image–point-cloud tiles, each paired with a corresponding segmentation mask. From the full dataset, 2970 image–point-cloud pairs were allocated for model development, which includes both the training and validation phases, while the remaining 384 pairs were set aside exclusively for testing. By isolating this test subset, we ensure that final performance metrics reflect unbiased model behavior, free from any influence of the training or validation processes. Table 1 summarizes the basic spatial characteristics of the areas utilized in this study.

Table 1. Detailed breakdown of the dataset used for our research.

Area Name	Area (km^2)	Image–Point Cloud Pairs Extracted	Usage
Area #1	5.05 km^2	2860	Train
Area #2	0.19 km^2	110	Train
Area #3	0.68 km^2	384	Test

2.2. Methodology

2.2.1. Point Cloud Sampling

Our methodology employs pixel-wise fusion of LiDAR and RGB imagery to achieve accurate semantic segmentation. The proposed method does not perform explicit point-to-point matching or image-recognition-based correspondence between RGB imagery and LiDAR data. Instead, both data sources are assumed to be georeferenced in a common spatial reference system, enabling a direct geometric projection of LiDAR points onto the image plane. Each LiDAR point is associated with its corresponding image pixel based solely on spatial coordinates, for which ground control points and standard aerial triangulation are sufficient to ensure reliable alignment. To ensure correspondence between points and pixels, it is essential to balance the resolution of both datasets. In this study, we employed 128 × 128-pixel RGB images for segmentation tasks. For the LiDAR data, we sampled 16,384 points from each sub-point cloud to ensure that all the points in the point cloud could be paired with the corresponding pixel. For the current density and number of points in the point cloud segments of our dataset, this number of sampled points secured the best quality of pixel-point matching, as only a small number of pixel segments were left without any points falling into their area. This approach ensures consistency in data representation across modalities, enabling effective, pixel-level fusion. Let $P = \{p_i\}_{i=1}^{M}$, $p_i = (X_i, Y_i, Z_i)$ be the LiDAR point cloud where X and Y are horizontal coordinates, and Z is elevation. Also, let the image pixel grid coordinates (u, v) with $u \in \{0, \ldots, W-1\}$, $v \in \{0, \ldots, H-1\}$, where $W = H = 128$. During the sampling process, we took each sub-area of the point cloud that covers the area of one pixel as

$$\pi(p_i) = (u,\ v)\ , \tag{1}$$

where $\pi(p)$ function determines the pixel coordinate location of the given point. By taking the top-left coordinate as (0, 0) of the point cloud area and measuring the distance of each

point from that reference mark, each p_i point can be assigned to a $(u,\ v)$ pixel coordinate where each pixel coordinate covers 30 cm $\times$ 30 cm of the point cloud. After the assignment of each point cloud coordinate, we can define a LiDAR point set for individual pixels u and v like,

$$S_{uv} = \{p_i \epsilon\ P :\ \pi(p_i) = (u,\ v)\}, \tag{2}$$

where S_{uv} is the set of p_i points, where the point in the point cloud is assigned to the pixel coordinate area defined by (u, v) pixel coordinates. Let I be the 2D image with the width of W and the height of H. The mapping of the 3D elevation values to 2D elevation values is done by taking the average of the elevation values of each point within the $S_{u,v}$ set of points.

$$I_{uv} = \frac{1}{|S_{u,v}|}\sum_{p \in s_{uv}} p_z, \tag{3}$$

where I_{uv} is a pixel on I image at (u, v) coordinates. $|S_{u,v}|$ is the number of points in the S set and p_z is the elevation value of the point p.

Figure 1 demonstrates the whole sampling process of how the sampled pixel is calculated for each sub-section of the point cloud.

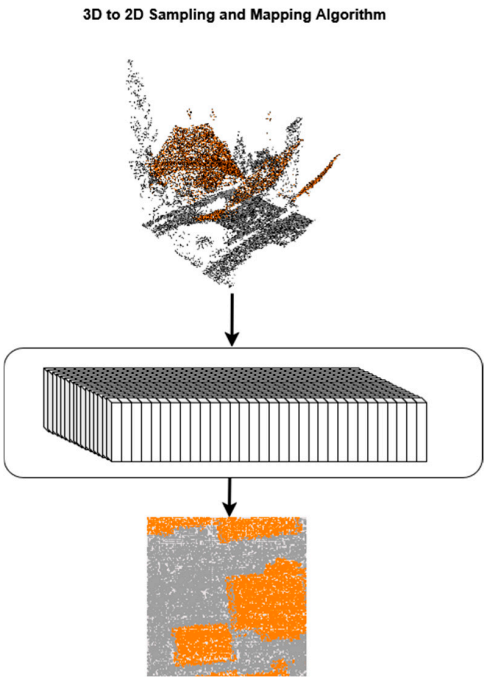

Figure 1. Representation of sampling algorithm converting a batch of 3D points into 2D pixels.

Although mean pooling may smooth local geometric variations, sharp building boundaries are primarily preserved through RGB feature extraction and multi-scale decoder fusion, as confirmed by the ablation results presented in Section 3.1. The number of LiDAR points contributing to each pixel is not fixed but is determined implicitly by the LiDAR point density and the pixel footprint on the ground; all points projected into a given pixel are aggregated without explicit subsampling.

2.2.2. Train with Merged RGB and Flattened Point Cloud

The sampling strategy can be integrated at different stages within the processing pipeline, and we explored several approaches to fuse LiDAR information with RGB imagery. In the initial approach, the sampling algorithm is deployed as a preprocessing step to establish a per-point correspondence between the LiDAR cloud and the RGB pixels. Specifically, we project the LiDAR points onto the image plane to assign each point to a

target pixel. We then derive a 2D height image by using the Z-coordinate values of the points, and normalize this height image to the [0, 1] range. As a final preprocessing step, the normalized height image is concatenated with the RGB image to form a 4-channel input for the semantic segmentation network. The overall preprocessing architecture is summarized in Figure 2.

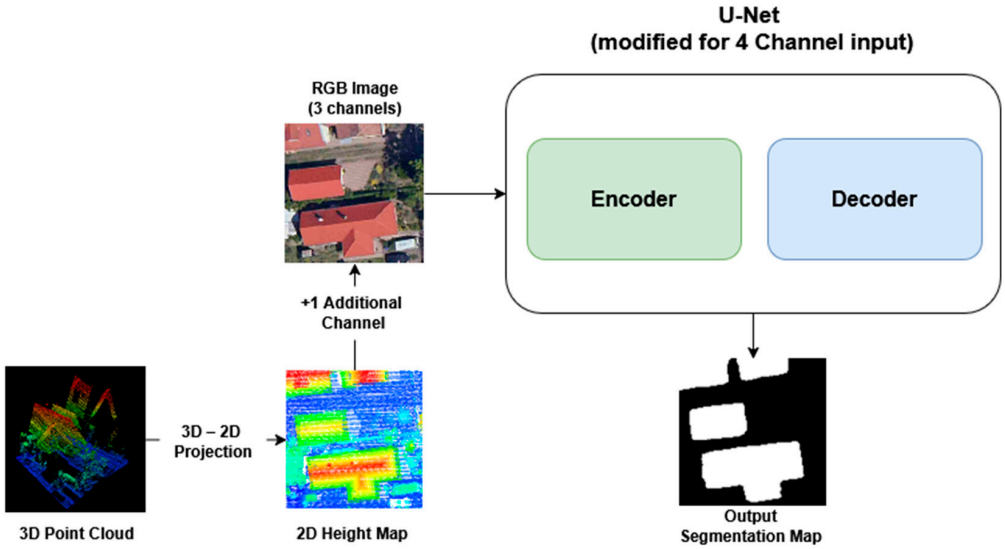

Figure 2. Architecture and pipeline of 4D U-Net segmentation network.

2.2.3. Point U-Net

Our principal contribution is the Point U-Net semantic segmentation architecture, which fuses geometric and spatial information provided by the LiDAR point cloud with RGB imagery by integrating two segmentation backbones: PointNet++ and a modified U-Net semantic segmentation model. This architecture supports flexible training strategies, including the use of pretrained weights for the point-cloud branch or simultaneous training of both branches in parallel. The high-level overview of the proposed architecture is depicted in Figure 3. This figure illustrates the principal components of the Point U-Net architecture as employed during training. The PointNet++ branch can be utilized as a pretrained model operating solely on point-cloud data, providing a strong baseline for mutual segmentation. Alternatively, each component of the Point U-Net can be trained jointly, incorporating both the PointNet++ branch and the auxiliary U-Net (O-Net) branch. As the initial phase of training, the 3D LiDAR point cloud is fed into the point-cloud segmentation branch, which is built upon the PointNet++ architecture.

Within this phase, the network extracts local geometric features from the point cloud by employing the Set Abstraction and Feature Propagation modules, yielding a 3D feature map with corresponding activation values. Just prior to the final segmentation step, these per-point activation values are extracted by the proposed method. The point cloud is then reshaped into a 2D image where pixel intensities correspond to the activation values, while the per-point coordinates from the original cloud are retained to preserve spatial correspondence.

The sampling strategy of our solution matches the pixels on the RGB image with a single point from the point cloud, we can be sure that the transformed points on the 2D map represent the matching pixel on the 2D image. As an additional input, during the preprocessing step, the point cloud is projected directly onto the 2D field to create a depth map of the point cloud which also enhances the segmentation. This depth map is merged into the 2D activation map of the point cloud to get a weighted activation map, which helps during the further segmentation steps of the semantic segmentation model.

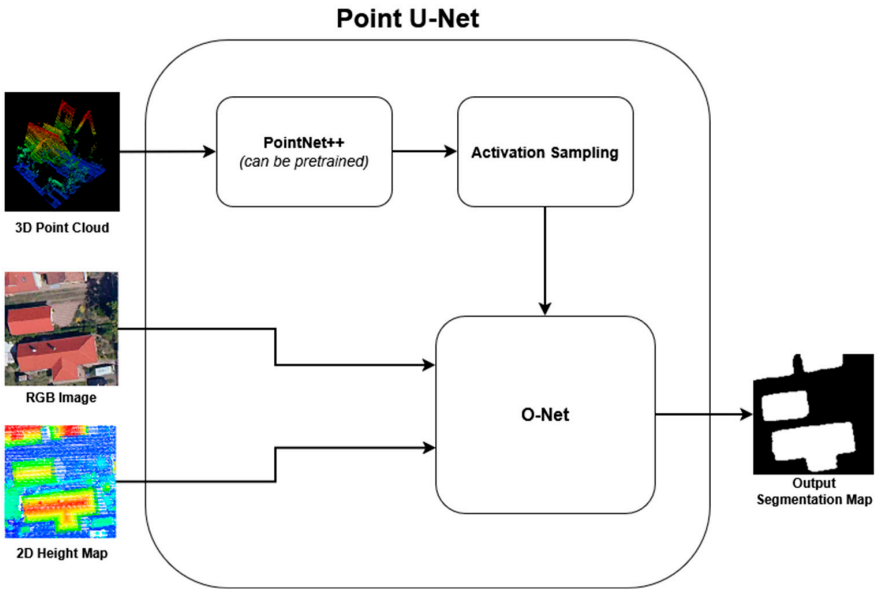

Figure 3. High-level flowchart of the proposed Point U-Net architecture with PointNet++, activation sampling algorithm and a modified U-Net (O-Net).

The second stage of the network takes the generated 2D map from the LiDAR point cloud projection and the corresponding 2D RGB aerial image as input. In this stage, our modified U-Net-based architecture is designed to process both inputs in parallel during the encoding phase, enabling the model to extract complementary features from each data source efficiently. This dual-input strategy allows the network to simultaneously capture detailed texture and color information from the RGB image and precise structural and elevation details from the LiDAR-derived 2D projection. Such an approach significantly enhances the overall feature representation before fusion, leading to improved segmentation accuracy and more precise delineation of building boundaries. A key advantage of our proposed architecture is its flexibility—the most suitable encoder can be selected for each input type, meaning that two distinct encoder networks can be employed for the RGB image and the 2D point cloud projection, respectively. The high-level representation of the proposed Point U-Net is illustrated in Figure 3.

This modularity allows the network to be adapted for different datasets, input resolutions, and computational constraints. During our experiments, we evaluated several state-of-the-art convolutional encoder backbones, including EfficientNetB3, EfficientNetB7, ResNet50, and MobileNetV2, each known for its unique balance between accuracy, efficiency, and computational cost. Through extensive testing and validation on our dataset, which consists of more than 3000 aerial images and their corresponding point clouds, we determined that the EfficientNetB3 encoder offered the best trade-off between segmentation performance and model complexity. The superior results achieved with EfficientNetB3 can be attributed to its compound scaling approach, which uniformly balances network depth, width, and resolution, allowing it to extract high-quality features without excessively increasing computational load. Furthermore, its architecture effectively captures both global and local contextual information, making it particularly well-suited for complex urban environments where buildings vary in shape, size, and texture. Once the encoding phase is complete, the extracted features from both modalities are progressively fused during the decoding phase, where the U-Net's skip connections ensure that spatial information is preserved throughout the reconstruction process.

As illustrated in Figure 4, the complete processing pipeline of the modified U-Net segmentation network begins with the preprocessing and projection of the LiDAR point cloud

into a 2D representation, followed by the parallel encoding of the RGB and LiDAR-derived inputs. The decoder then merges these learned features to produce a high-resolution segmentation map that accurately identifies building footprints. This comprehensive design not only enhances segmentation accuracy but also demonstrates the scalability and adaptability of multimodal fusion networks for remote sensing and geospatial analysis tasks.

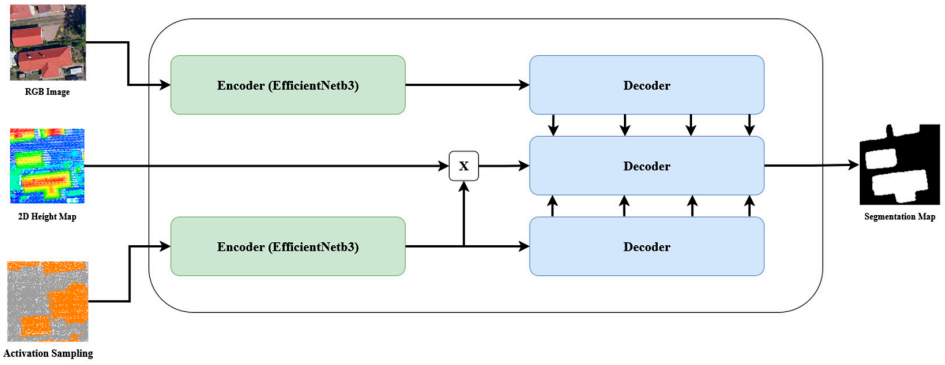

Figure 4. Point U-Net architecture with 3 inputs (RGB image, 2D Height Map, Projected 3D activation points).

The decoder phase of the network is responsible for integrating the information extracted from both the RGB image and the LiDAR-derived 2D projection, enabling the model to generate a unified and detailed segmentation map. During each stage of decoding, the network merges the two parallel input streams to progressively reconstruct the spatial structure of the scene. The merging process begins with the concatenation of the current feature maps from the RGB branch and the projected activation values derived from the point cloud branch. This step ensures that both spectral–textural and geometric–structural features are combined at every decoding level, providing a richer feature representation for the subsequent layers. After concatenation, the combined feature maps are passed through a double convolutional block designed to refine the fused features. This block consists of two consecutive 2D convolutional layers, each followed by a Batch Normalization layer and a ReLU activation function.

The use of double convolutional layers enhances the network's ability to capture both fine-grained and high-level contextual information while maintaining training stability and efficient gradient propagation through normalization. The architecture preserves separate decoding paths for both the RGB and point cloud branches. These independent decoders allow each modality to retain its unique feature hierarchies, ensuring that modality-specific information is not lost during the fusion process. As a result, the RGB decoder continues to focus on texture, color, and appearance-based cues, while the point cloud decoder emphasizes depth, elevation, and geometric consistency. The merging path then draws complementary insights from both, leading to more robust and context-aware segmentation outcomes.

Unlike standard U-Net architectures, where feature fusion is performed solely through skip connections, the proposed Point U-Net employs a multi-branch decoder that processes RGB features, LiDAR-derived geometric features, and their fused representations in parallel. Let ${F_r}^n$ denote the RGB feature map and ${F_l}^n$ the projected LiDAR feature map at decoder level n. At each decoder block, modality-specific features are first refined independently and subsequently combined through channel-wise concatenation followed by convolution (4):

$$ {F_f}^n = \sigma({F_f}^{n-1} \oplus [{W_f}^n * \{{F_r}^n \oplus {F_l}^n\}]), \quad (4) $$

where $\oplus$ represents the channel-wise concatenation, ${W_f}^n$ is a learnable convolutional kernel at level n, ${F_f}^{n-1}$ is the fused layer at level $n-1$ and $\sigma(\cdot)$ is a non-linear activation function.

This decoder-level fusion strategy allows geometric cues from LiDAR data to influence boundary refinement without suppressing high-frequency texture information from the RGB modality, which distinguishes the proposed architecture from conventional U-Net and PointNet++ designs.

As illustrated in Figure 5, the full decoder structure demonstrates how each modality contributes distinct yet synergistic information to the final segmentation map. At the concluding stage of the decoding process, the outputs of all three branches—the RGB decoder, the point cloud decoder, and the merging path—are concatenated to form a comprehensive feature representation. A final 2D convolutional layer then processes this aggregated feature set, representing the predicted segmentation output of the network.

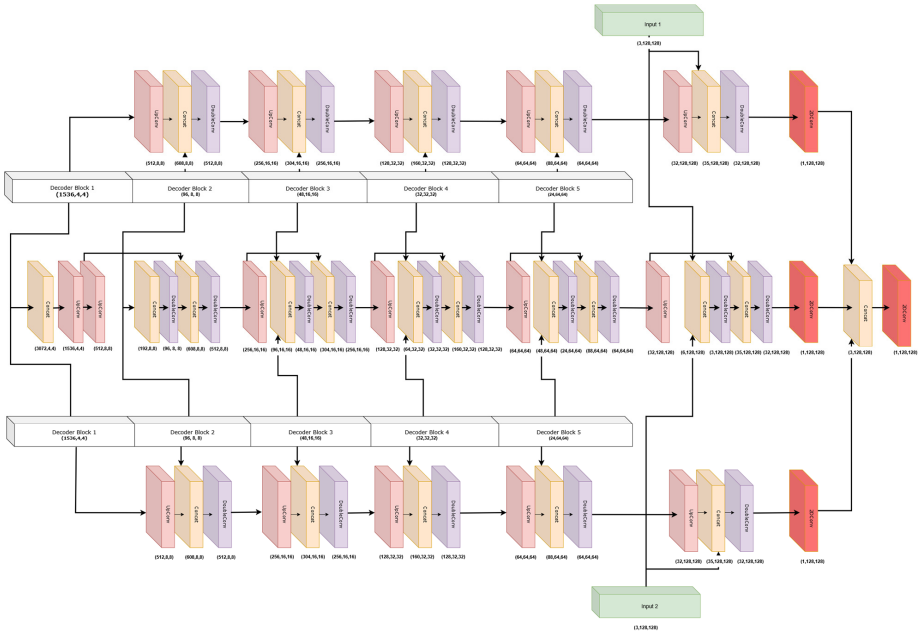

Figure 5. Blocks of the O-Net decoders.

The previous figures illustrated the overall structure and the detailed data flow of the proposed architecture; Table 2 provides a structured overview of the encoder, projection, and decoder modules, including modality-specific branches, feature dimensionalities, and spatial resolutions.

Table 2. Structural overview of the proposed Point-U-Net architecture.

Module	Modality	Stage	Input	Operation	Output Resolution	Feature Dimension
Encoder	RGB Image	E1	RGB Image	EfficientNet-B3 stem + MBConv blocks	128 × 128	40
		E2	E1	MBConv blocks	64 × 64	48
		E3	E2	MBConv blocks	32 × 32	136
		E4	E3	MBConv blocks	16 × 16	384
Encoder	Point Cloud	L1	Point Cloud	PointNet++ set abstraction	Point-based	128
		L2	L1	PointNet++ set abstraction	Point-based	256
		L3	L2	PointNet++ set abstraction	Point-based	512
Projection	Point Cloud	P	L3	Point-to-pixel projection with mean pooling	128 × 128	1
Auxiliary Input	Point Cloud	D	Point Cloud	Rasterized height (depth) map	128 × 128	1

Table 2. *Cont.*

Module	Modality	Stage	Input	Operation	Output Resolution	Feature Dimension
Decoder	RGB Image	RD-1	Bottleneck + skip (E4)	Upsampling + Conv	8 × 8	512
		RD-2	RD-1 + skip (E3)	Upsampling + Conv	16 × 16	256
		RD-3	RD-2 + skip (E2)	Upsampling + Conv	32 × 32	128
		RD-4	RD-3 + skip (E1)	Upsampling + Conv	64 × 64	64
		RD-5	RD-4	Upsampling + Conv	128 × 128	32
		RD-6	RD-5 + RGB Image	Upsampling + Concatenation + Conv	128 × 128	32
	Point Cloud	PD-1	Projected Point Features	Upsampling + Conv	8 × 8	512
		PD-2	PD-1	Upsampling + Conv	16 × 16	256
		PD-3	PD-2	Upsampling + Conv	32 × 32	128
		PD-4	PD-3	Upsampling + Conv	64 × 64	64
		PD-5	PD-4	Upsampling + Conv	128 × 128	32
		PD-6	PD-5 + Projected Point Cloud	Upsampling + Concatenation + Conv	128 × 128	32
	Fusion	FD-1	PD-1 + RD-1	Concatenate + Upsampling + Conv	4 × 4	512
		FD-2	PD-2 + RD-2 + FD-1	Concatenate + Upsampling + Conv	8 × 8	512
		FD-3	PD-3 + RD-3 + FD-2	Concatenate + Upsampling + Conv	16 × 16	256
		FD-4	PD-4 + RD-4 + FD-3	Concatenate + Upsampling + Conv	32 × 32	128
		D-5	PD-5 + RD-5 + FD-4	Concatenate + Upsampling + Conv	64 × 64	64
		FD-6	PD-6 + RD-6 + FD-5	Concatenate + Upsampling + Conv	128 × 128	32
Final Fusion	Multimodal	F	RD-6 + PD-6 + FD-6	Concatenation + Conv	128 × 128	1

3. Results

In this section, we present a comprehensive experimental evaluation of the proposed Point U-Net semantic segmentation network. The evaluation is structured in sub-sections. First, an ablation study is conducted to systematically analyze the contribution of the key design choices of the proposed method, including the LiDAR feature sampling strategy, the employed fusion mechanism, and the multi-branch decoder architecture. This analysis aims to quantify the individual impact of each component on segmentation performance. Second, the full model is compared against unimodal baselines and existing multimodal approaches to assess its effectiveness under identical experimental conditions. For training, we used two of our three available areas, and we performed the evaluation of the semantic segmentation network in the third area, dedicated only for testing purposes. To measure the performance of the architecture, we used the *Jaccard* score [38] and calculated it for each result per input pair as

$$Jaccard(gt, P) = \frac{|gt \cap P|}{|gt \cup P|}, \tag{5}$$

where gt stands for the ground truth and P for the actual prediction of the semantic segmentation model. Similar to the *Jaccard* score, we also evaluate the different semantic segmentation methods by calculating the *Dice* coefficient as shown below:

$$Dice(gt, P) = \frac{2|gt \cap P|}{|gt| + |P|}, \tag{6}$$

Precision and *Recall* provide a more detailed analysis of segmentation performance. *Precision* measures the proportion of correctly predicted building pixels among all predicted

building pixels, while *Recall* measures the proportion of correctly detected building pixels relative to the ground truth. These metrics are defined as:

$$Precision = \frac{TP}{TP + FP}, \tag{7}$$

$$Recall = \frac{TP}{TP + FN}, \tag{8}$$

where TP denotes the true positive, FP denotes the false positive and FN the false negative predictions. We also evaluated boundary accuracy using the Boundary Intersection over Union (*Boundary IoU*) [39], which focuses on the agreement between predicted and ground truth object contours.

$$BoundaryIoU(gt, P) = \frac{|B(gt) \cap B(P)|}{|B(gt) \cup B(P)|}, \tag{9}$$

where $B(gt)$ and $B(p)$ denote the boundary regions of the ground truth and predicted masks, respectively, defined as pixels within a fixed distance δ from the object boundaries. The boundary distance $\delta = 5$ for our evaluation.

3.1. Ablation Studies

To better understand the contribution of the individual components of the proposed Point U-Net framework, a series of controlled ablation experiments was conducted. All ablated variants were trained and evaluated using identical dataset splits, training schedules, and evaluation protocols as the full model, with performance assessed on the independent test area. Unless otherwise stated, only a single architectural component was modified at a time to isolate its impact on the final building footprint segmentation performance.

3.1.1. Effect of LiDAR Feature Sampling Strategy

The proposed Point U-Net architecture employs mean pooling for projecting the 3D point cloud features onto a 2D elevation map. Mean pooling was selected to improve robustness against noise and local point-density variations commonly present in airborne LiDAR data. However, alternative aggregation strategies such as max pooling and min pooling are also frequently adopted in point-based fusion approaches. To evaluate the influence of the feature aggregation strategy, we replaced the mean pooling operation with max and min pooling while keeping all other network components unchanged. which is presented in Table 3.

Table 3. Performance comparison of different components of the proposed architecture on our test dataset. All variants are compared to the final architecture; see the performance difference in the brackets for all metrics.

Category	Variant	Precision	Recall	IoU	Dice	Boundary IoU
Sampling	Max pooling	0.7985 ± 0.0175 (−0.0625)	0.8229 ± 0.0218 (−0.0860)	0.7821 ± 0.0156 (−0.0466)	0.8093 ± 0.0195 (−0.0576)	0.5996 ± 0.0122 (−0.0332)
	Min pooling	0.7905 ± 0.0245 (−0.0705)	0.8301 ± 0.0146 (−0.0788)	0.7891 ± 0.0196 (−0.0396)	0.8106 ± 0.0213 (−0.0563)	0.6069 ± 0.0184 (−0.0259)
Fusion	Without Depth Map	0.8096 ± 0.0194 (−0.0514)	0.8163 ± 0.0219 (−0.0926)	0.7915 ± 0.0203 (−0.0372)	0.8111 ± 0.0233 (−0.0558)	0.6008 ± 0.0257 (−0.0320)
Decoder	Single Decoder	0.8055 ± 0.0187 (−0.0555)	0.8194 ± 0.0275 (−0.0895)	0.7701 ± 0.0269 (−0.0586)	0.7898 ± 0.0291 (−0.0771)	0.5914 ± 0.0290 (−0.0414)
	Late Fusion Decoder	0.7522 ± 0.0307 (−0.1088)	0.7776 ± 0.0399 (−0.1313)	0.7105 ± 0.0359 (−0.1182)	0.7490 ± 0.0408 (−0.1179)	0.4811 ± 0.0302 (−0.1517)

3.1.2. Contribution of Multimodal Fusion Components

To quantify the contribution of the individual multimodal fusion components, multiple reduced versions of the proposed architecture were evaluated by selectively disabling specific inputs while keeping the network topology unchanged. The sampling strategy of the point cloud was already discussed in the previous section. The Point U-Net framework incorporates complementary information sources on top of the original fusion strategy of the two modalities, introducing the depth map of the point cloud. Table 3 summarizes how the performance of our proposed Point U-Net network varies if this component is removed from the architecture.

3.1.3. Decoder Architecture Analysis

The current architecture of the Point U-Net has a three-branch decoder which channels the encoded point cloud activation map, the fusion of the depth map and activation map of the point cloud and the RGB image separately. To assess the effectiveness of this design choice, two alternative decoder configurations were evaluated. In the first variant, all modality-specific features were fused prior to the decoding stage and processed using a single decoder branch. In the second variant, the intermediate fusion branch was removed, allowing the RGB and point-cloud decoders to operate independently, with feature fusion performed only at the final decoding layer. The results of these experiments are presented in Table 3 above.

3.2. Comparative Performance Evaluation

To compare our proposed architecture with existing solutions, we trained and tested all models on the same dataset using identical hyperparameters, ensuring a fair and consistent evaluation. The training was conducted with a batch size of 16 and a learning rate of 0.001, while the evaluation measured the mean segmentation accuracy and included variance to assess model stability. The models were trained using the Adam optimizer and IoU loss with an early stopping strategy based on the validation loss to avoid overfitting on the training dataset. The models were not trained for more than 50 epochs; convergence was reached around 30–35 epochs in each case. All experiments were conducted on a workstation equipped with an NVIDIA A100 GPU (40 GB VRAM), an Intel multi-core CPU, and 128 GB of system memory. Training a single model required approximately 3–4 h, depending on the specific architecture variant. To ensure a comprehensive comparison, we included both RGB-only segmentation networks and point cloud-only segmentation networks, each tested with different encoder backbones where applicable. Table 4 summarizes the performance metrics of all evaluated models. As the results indicate, the RGB-based networks achieved comparable performance levels, as they primarily rely on visual feature extraction from the input imagery.

Table 4 summarizes all evaluated runs for each semantic segmentation model. As the results demonstrate, the RGB segmentation networks achieved equal performance, as they tend to have similar capabilities for extracting visual features from the input image. For the point cloud segmentation, the PointNet++ architecture proved to be the outstanding model for our dataset, as it significantly outperformed all other models designed for point set segmentation. For experimental purposes, we also tested a traditional ensemble method by averaging the output predictions of the two best-performing unimodal models, DeeplabV3+ and PointNet++. As per the prerequisites of any ensemble technique, we performed the same projection steps for the point cloud prediction mask, but only as a post-processing step. As the table shows, it did not provide the desired result, as the interpolation of the projected point cloud worsened the overall segmentation performance. To further investigate this limitation, we examined whether alternative fusion strategies—such as

feature-level or confidence-weighted ensembling—could mitigate the degradation, but preliminary tests indicated similar drawbacks. This suggests that naive post hoc combination of heterogeneous modalities cannot fully exploit their complementary strengths without more sophisticated alignment techniques. Additionally, the projection-induced artifacts highlight the sensitivity of multimodal integration to spatial consistency between RGB and 3D representations. Future work may therefore require learning-based fusion approaches that jointly optimize cross-modal representations instead of relying on handcrafted projection pipelines. The proposed Point U-Net was also compared with some current state-of-the-art multimodal RGB-LiDAR segmentation networks (DLAFNet and PMNet). For a fair comparison, these models were trained and evaluated using the same dataset splits and evaluation protocol as the proposed Point U-Net. When architectural constraints required differing hyperparameters, the recommended settings from the original implementations were adopted. Nonetheless, these experiments provide valuable insight into the challenges of integrating point cloud and image-based segmentation models in real-world scenarios.

Table 4. Quantitative comparison of the proposed Point U-Net with unimodal baselines and state-of-the-art multimodal RGB–LiDAR segmentation methods on the test area.

Model	Precision	Recall	Jaccard Score	Dice Coefficient	Boundary IoU Score
Only RGB Input					
U-Net (EfficientNet-b3)	0.8120 ± 0.0236	0.8889 ± 0.0178	0.7365 ± 0.0258	0.7860 ± 0.0411	0.5201 ± 0.0378
U-Net (ResNet50)	0.8105 ± 0.0205	0.8814 ± 0.0194	0.7315 ± 0.0193	0.7683 ± 0.0358	0.5155 ± 0.0272
U-Net (MobileNetV2)	0.7963 ± 0.0245	0.8743 ± 0.0258	0.7049 ± 0.0154	0.7446 ± 0.0301	0.5299 ± 0.0165
PSPNet	0.7625 ± 0.0366	0.8225 ± 0.0359	0.6398 ± 0.0311	0.7244 ± 0.0452	0.5051 ± 0.0049
DeeplabV3+	0.8129 ± 0.0211	0.8901 ± 0.0257	0.7369 ± 0.0202	0.7962 ± 0.0374	0.5458 ± 0.0278
Only Point Cloud Input					
PointNet++	0.8369 ± 0.0287	0.8340 ± 0.0243	0.8047 ± 0.0299	-	-
RandLANet	0.6488 ± 0.0589	0.6626 ± 0.0497	0.6022 ± 0.0689	-	-
Ensemble Methods					
Average method	0.7514 ± 0.0402	0.7701 ± 0.0394	0.6128 ± 0.0302	0.6957 ± 0.0448	0.4516 ± 0.0207
Multimodal solutions					
4D RGB + Flattened Point Cloud	0.8384 ± 0.0277	0.8563 ± 0.0230	0.7602 ± 0.0298	0.8162 ± 0.0421	0.4733 ± 0.0403
DLAFNet	0.7952 ± 0.0255	0.8032 ± 0.0280	0.7756 ± 0.0304	0.7965 ± 0.0244	0.5620 ± 0.0266
PMNet	0.7801 ± 0.0355	0.7993 ± 0.0391	0.7601 ± 0.0346	0.7901 ± 0.0415	0.5432 ± 0.0487
Our Solution (with pretrained PointNet)	**0.8501 ± 0.0168**	**0.8814 ± 0.0201**	**0.8179 ± 0.0202**	**0.8341 ± 0.0376**	**0.6301 ± 0.0102**
Our Solution (without pretrained PointNet)	**0.8610 ± 0.0151**	**0.9089 ± 0.0199**	**0.8287 ± 0.0169**	**0.8669 ± 0.0306**	**0.6328 ± 0.0117**

The main improvement of the multimodal segmentation is the accuracy of detecting the edges of the objects, as the point cloud input brings the necessary spatial information about the buildings, which, combined with the visual representation from the RGB images, enhances the segmentation performance significantly. To assess the statistical significance of the observed performance differences, a paired *t*-test was conducted between the proposed Point U-Net and the PointNet++ baseline across all test samples. The results indicate that the improvement in IoU is statistically significant at the 95% confidence level ($p < 0.05$), confirming that the observed gain is unlikely to be due to random variation. Figure 6 visually demonstrates how the 3D information gained from the point clouds enhances the segmentation accuracy in different scenarios. This improvement is especially noticeable in regions where RGB features alone are ambiguous or affected by shadows and varying illumination conditions. Complementary depth cues help the model better delineate object boundaries and reduce the occurrence of false positives along complex structural edges. As a result, the multimodal approach produces more robust and spatially coherent predictions across diverse urban environments.

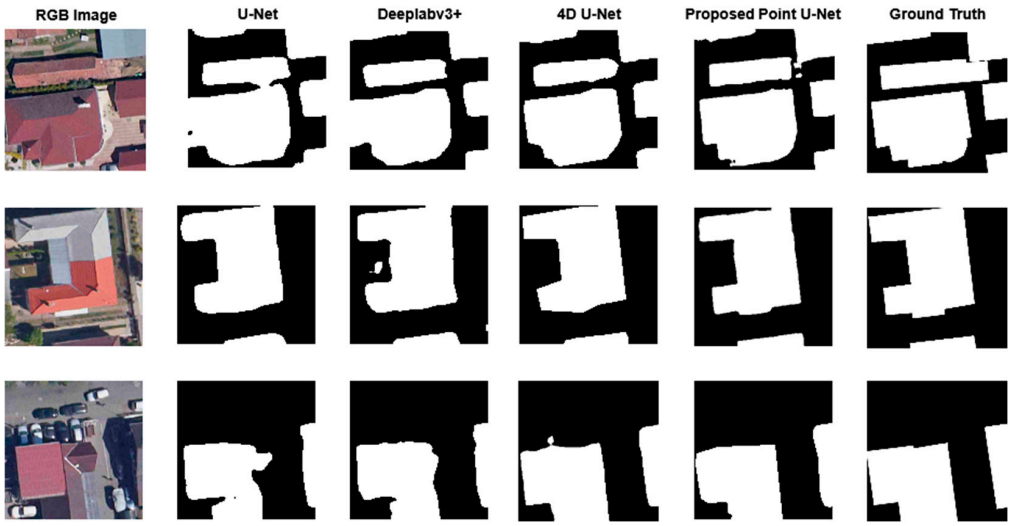

Figure 6. Visual comparison of different semantic segmentation methods on the test dataset.

To further evaluate segmentation performance at the pixel level, Figure 7 presents a detailed visual comparison highlighting correct and incorrect predictions for each method. In this visualization, correctly classified building pixels are shown in green, while mismatches relative to the ground truth are marked in red, providing a fine-grained view of prediction accuracy, particularly along building boundaries and in complex urban structures. As shown in Figure 7, the proposed Point U-Net exhibits fewer mismatched regions and improved boundary consistency compared to the baseline models.

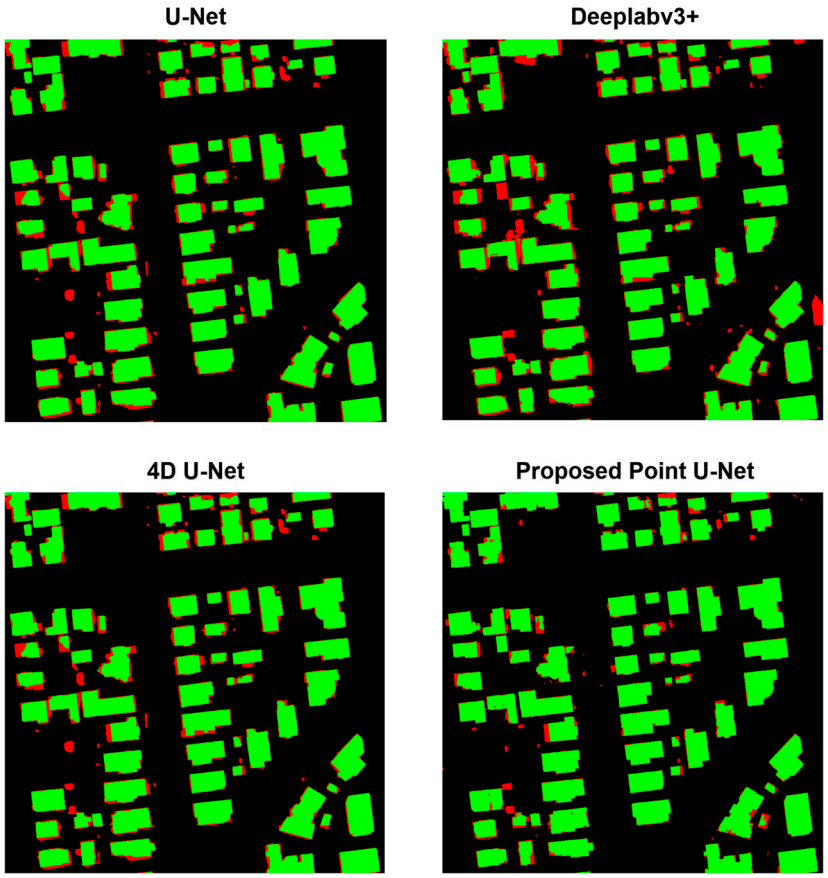

Figure 7. Pixel-wise comparison of segmentation performance across models. Green pixels indicate correct predictions, while red pixels represent mismatches with the ground truth.

4. Discussion

The experimental results confirm that the new multimodal model integrating PointNet++ with a modified U-Net architecture (Point U-Net) outperforms traditional single-modality approaches based on either point cloud or image data. This improvement demonstrates the complementarity of 3D geometric information and 2D contextual cues. While image-based segmentation has rich spatial resolution and texture, it is prone to occlusion, lighting variation, and scale variation. In contrast, point cloud segmentation is well-equipped to handle geometric structure and spatial relationships but suffers from sparsity and irregular sampling. By using both modalities within a unified framework, the proposed method avoids the respective disadvantages of individual representations. The proposed method also performs better than DLAFNet and PMNet based on our quantitative measures under identical evaluation conditions, indicating that point-level feature sampling and multi-branch decoding provide measurable benefits over existing multimodal fusion strategies. However, despite the observed performance gains, it is worth noting that PMNet and DLAFNet were originally designed for different dataset characteristics, and their performance may vary under alternative acquisition settings or higher-resolution LiDAR point densities.

The most significant outcome of the multimodal fusion is the sustained improvement in performance across challenging cases such as object boundaries, thin lines, and regions with sparse depth information. These gains suggest that feature interactions between modalities learned by the architecture result in more robust representations than modality-specific baselines. Further, PointNet++'s hierarchical feature abstraction is closely in agreement with U-Net's encoder–decoder structure, allowing for efficient multi-scale fusion without excessive computational overhead.

Although the RGB images are downscaled to 128 $\times$ 128 pixels to ensure reliable point-to-pixel correspondence with LiDAR data, this does not imply that the achievable building footprint accuracy is limited to the resulting pixel spacing. The downsampling step is a pragmatic design choice that stabilizes multimodal fusion by avoiding empty or sparsely supported pixels, while fine-grained boundary information is preserved through RGB-based feature extraction, multi-scale decoding, and LiDAR-derived geometric cues. Consequently, building footprint quality is determined by the learned feature representations and boundary consistency rather than by the nominal pixel size alone. This design trade-off prioritizes geometric consistency and cross-modal alignment over raw spatial sampling density, which we found to be essential for robust footprint extraction in multimodal settings.

However, several limitations should be mentioned. First, the application of paired and well-registered multimodal data introduces a dependency that can restrict generalization of the model to datasets in which both modalities are consistently present and properly registered. Although the point-level feature sampling strategy and the decoder-level multimodal fusion introduce a degree of spatial tolerance, the method does not explicitly correct geometric misalignment between RGB imagery and LiDAR data. Small registration offsets can be attenuated through neighborhood aggregation and complementary modality processing; however, larger systematic misalignments—for example, those caused by orthorectification inaccuracies or sensor calibration errors—may still negatively affect the precise delineation of building boundaries. Second, the proposed approach relies on rasterized representations and local sampling strategies that implicitly assume a sufficient local point density. In areas with extremely sparse LiDAR coverage or highly heterogeneous point distributions, the benefit of point-level sampling may be reduced. Moreover, the model performs with enhanced precision, but it comes at the cost of increased training complexity and memory consumption over unimodal models, which could restrict its deployment in environments with strict resource budgets. The scope of the experimental

evaluation should also be considered when interpreting the results. All experiments were conducted within a single geographic region under a consistent sensing configuration, which was intentionally selected to provide a controlled setting for analyzing the impact of the proposed multimodal fusion strategy. Although this allows for a rigorous architectural evaluation, differences in urban characteristics, sensor configurations, or acquisition conditions may affect performance in other scenarios. While the core design of the proposed method is expected to be broadly applicable, additional validation across heterogeneous datasets and geographic regions is required to fully assess its generalization capability and remains a topic for future work.

Overall, the results indicate multimodal fusion with the proposed PointNet++ and U-Net hybrid architecture is a promising avenue for semantic segmentation. Closing the gap between geometric reasoning and visual context, the approach not only surpasses unimodal baselines but also provides an extendible point of departure for extensions such as attention-based fusion, transformer layers, or domain adaptation across vastly different datasets.

5. Conclusions

In summary, this work presented a multimodal semantic segmentation framework that integrates PointNet++ with a modified U-Net, effectively combining 3D geometric information with 2D visual context. The proposed architecture consistently outperformed unimodal baselines, particularly in complex regions where either images or point clouds alone were insufficient. While challenges remain regarding computational cost and dependence on accurately registered multimodal data, the results confirm the potential of cross-modal fusion for advancing semantic segmentation. Future directions include optimizing the fusion strategy, extending the dataset of other regions for enhanced generalization, and exploring lightweight variants for real-time applications.

Author Contributions: Conceptualization, N.S. and B.H.; Methodology, N.S.; Software, N.S.; Validation, B.H., P.E. and P.B.; Formal Analysis, N.S.; Investigation, N.S.; Resources, P.E. and P.B.; Data Curation, P.E. and P.B.; Writing—Original Draft Preparation, N.S.; Writing—Review and Editing, N.S. and B.H.; Visualization, N.S.; Supervision, B.H.; Project Administration, B.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from Envirosense Hungary Ltd. and are available from the authors with the permission of Envirosense Hungary Ltd.

Conflicts of Interest: Péter Enyedi is affiliated with Envirosense Hungary Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Nurkarim, W.; Wijayanto, A.W. Building footprint extraction and counting on very high-resolution satellite imagery using object detection deep learning framework. *Earth Sci. Inform.* **2023**, *16*, 515–532.
2. Mharzi Alaoui, H.; Radoine, H.; Chenal, J.; Hajji, H.; Yakubu, H. Deep building footprint extraction for urban risk assessment–Remote sensing and Deep learning based approach. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *48*, 83–86. [CrossRef]
3. Shan, J.; Toth, C.K. (Eds.) *Topographic Laser Ranging and Scanning: Principles and Processing*; CRC Press: Boca Raton, FL, USA, 2018.
4. Ritchie, J.C.; Everitt, J.H.; Escobar, D.E.; Jackson, T.J.; Davis, M.R. Airborne laser measurements of rangeland canopy cover and distribution. *J. Range Manag.* **1992**, *45*, 189–193. [CrossRef]
5. Chen, S.; Shi, W.; Zhou, M.; Zhang, M.; Chen, P. Automatic building extraction via adaptive iterative segmentation with LiDAR data and high spatial resolution imagery fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 2081–2095. [CrossRef]

6. Liang-chien, C.; Tee-ann, T.; Yi-chen, S.; Yen-chung, L.; Jiann-yeou, R. Fusion of Lidar Data and Optical Imagery for Building Modeling. In *Geo-Imagery Bridging Continents XXth ISPRS Congress*; International Society for Photogrammetry and Remote Sensing: Hannover, Germany, 2004; pp. 12–23.
7. Matikainen, L.; Hyyppä, J.; Hyyppä, H. Automatic detection of buildings from laser scanner data for map updating. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2003**, *34*, 218–224.
8. Zhang, W.; Qi, J.; Wan, P.; Wang, H.; Xie, D.; Wang, X.; Yan, G. An easy-to-use airborne LiDAR data filtering method based on cloth simulation. *Remote Sens.* **2016**, *8*, 501. [CrossRef]
9. Nex, F.; Remondino, F. UAV for 3D mapping applications: A review. *Appl. Geomat.* **2014**, *6*, 1–15.
10. Jozi, D.; Shirzad-Ghaleroudkhani, N.; Luhadia, G.; Abtahi, S.; Gül, M. Rapid post-disaster assessment of residential buildings using Unmanned Aerial Vehicles. *Int. J. Disaster Risk Reduct.* **2024**, *111*, 104707. [CrossRef]
11. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
12. Khan, B.A.; Jung, J.-W. Semantic segmentation of aerial imagery using u-net with self-attention and separable convolutions. *Appl. Sci.* **2024**, *14*, 3712. [CrossRef]
13. Dechesne, C.; Lassalle, P.; Lefèvre, S. Bayesian u-net: Estimating uncertainty in semantic segmentation of earth observation images. *Remote Sens.* **2021**, *13*, 3836. [CrossRef]
14. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2018.
15. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
16. Azimi, S.M.; Fischer, P.; Korner, M.; Reinartz, P. Aerial LaneNet: Lane-marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 2920–2938.
17. Anilkumar, P.; Venugopal, P.; Maddikunta, P.K.R.; Gadekallu, T.R.; Al-Rasheed, A.; Abbas, M.; Soufiene, B.O. An adaptive DeepLabv3+ for semantic segmentation of aerial images using improved golden eagle optimization algorithm. *IEEE Access* **2023**, *11*, 106688–106705. [CrossRef]
18. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
19. Zhao, J.; Li, Z.; Lei, Y.; Huang, L. Application of UAV RGB images and improved PSPNet network to the identification of wheat lodging areas. *Agronomy* **2023**, *13*, 1309. [CrossRef]
20. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef] [PubMed]
21. Hanyu, T.; Yamazaki, K.; Tran, M.; McCann, R.A.; Liao, H.; Rainwater, C.; Adkins, M.; Cothren, J.; Le, N. Aerialformer: Multi-resolution transformer for aerial image segmentation. *Remote Sens.* **2024**, *16*, 2930. [CrossRef]
22. Deng, G.; Wu, Z.; Xu, M.; Wang, C.; Wang, Z.; Lu, Z. Crisscross-global vision transformers model for very high resolution aerial image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–19. [CrossRef]
23. Li, X.; Cheng, Y.; Fang, Y.; Liang, H.; Xu, S. 2dsegformer: 2-d transformer model for semantic segmentation on aerial images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
24. Zhang, J.; Zhao, X.; Chen, Z.; Lu, Z. A review of deep learning-based semantic segmentation for point cloud. *IEEE Access* **2019**, *7*, 179118–179133. [CrossRef]
25. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
26. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
27. Wang, G.; Wang, L.; Wu, S.; Zu, S.; Song, B. Semantic segmentation of transmission corridor 3D point clouds based on CA-PointNet++. *Electronics* **2023**, *12*, 2829. [CrossRef]
28. Hu, H.; Tan, Q.; Kang, R.; Wu, Y.; Liu, H.; Wang, B. Building extraction from oblique photogrammetry point clouds based on PointNet++ with attention mechanism. *Photogramm. Rec.* **2024**, *39*, 141–156. [CrossRef]
29. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. Randla-net: Efficient semantic segmentation of large-scale point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
30. Hertzberg, J. Hyperspectral 3D Point Cloud Segmentation Using RandLA-Net. In *Intelligent Autonomous Systems 17: Proceedings of the 17th International Conference IAS-17*; Springer Nature: Berlin/Heidelberg, Germany, 2023; Volume 577.

31. Kaijaluoto, R.; Kukko, A.; El Issaoui, A.; Hyyppä, J.; Kaartinen, H. Semantic segmentation of point cloud data using raw laser scanner measurements and deep neural networks. *ISPRS Open J. Photogramm. Remote Sens.* **2022**, *3*, 100011. [CrossRef]
32. Fang, X.; Wu, J.; Jiang, P.; Liu, K.; Wang, X.; Zhang, S.; Wang, C.; Li, H.; Lai, Y. A Rapid Assessment Method for Flood Risk Mapping Integrating Aerial Point Clouds and Deep Learning. *Water Resour. Manag.* **2024**, *38*, 1753–1772. [CrossRef]
33. Thomas, H.; Qi, C.R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; Guibas, L. Kpconv: Flexible and deformable convolution for point clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
34. Poliyapram, V.; Wang, W.; Nakamura, R. A point-wise LiDAR and image multimodal fusion network (PMNet) for aerial point cloud 3D semantic segmentation. *Remote Sens.* **2019**, *11*, 2961. [CrossRef]
35. Liu, W.; Wang, H.; Qiao, Y.; Zhang, H.; Yang, J. DLAFNet: Direct LiDAR-Aerial Fusion Network for Semantic Segmentation of 2D Aerial Image and 3D LiDAR Point Cloud. In *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*; IEEE: New York, NY, USA, 2024.
36. Yan, L.; Huang, J.; Xie, H.; Wei, P.; Gao, Z. Efficient depth fusion transformer for aerial image semantic segmentation. *Remote Sens.* **2022**, *14*, 1294. [CrossRef]
37. Wang, Y.; Wan, Y.; Zhang, Y.; Zhang, B.; Gao, Z. Imbalance knowledge-driven multi-modal network for land-cover semantic segmentation using aerial images and LiDAR point clouds. *ISPRS J. Photogramm. Remote Sens.* **2023**, *202*, 385–404. [CrossRef]
38. Jaccard, P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bull. Soc. Vaudoise Sci. Nat.* **1901**, *37*, 241–272.
39. Cheng, B.; Girshick, R.; Dollár, P.; Berg, A.C.; Kirillov, A. Boundary IoU: Improving object-centric image segmentation evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2021; pp. 15334–15342.