Height Estimation from Single Optical Images Using KANU-Net Architecture

Reyhaneh Vahabi[1*], Hossein Arefi[2], Reza Bahmanyar[3]

[1] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Iran – reyhane.vahabi@ut.ac.ir
[2] i3mainz, Institute for Spatial Information and Surveying Technology, School of Technology, Mainz University of Applied Sciences, D-55118 Mainz, Germany – hossein.arefi@hs-mainz.de
[3] Remote Sensing Technology Institute, German Aerospace Center, Wessling, Germany – reza.bahmanyar@dlr.de

Keywords: Deep Learning, KANU-Net, Digital Elevation Model, Height Estimation, Google Imagery, Urban Analysis.

Abstract

Monocular height estimation from single optical images is important for urban mapping and remote sensing, but remains challenging in heterogeneous urban scenes. We introduce KANU-Net, a U-Net variant that integrates Kolmogorov–Arnold Network (KAN) layers, which use functional basis expansions to enrich feature representation. KANU-Net is designed to better capture complex spatial patterns and multi-scale structures in aerial imagery. The method was evaluated on high-resolution (1 m) optical imagery from two urban areas: Utrecht (Google imagery) and Potsdam (ISPRS benchmark). Input data were processed into 256×256 patches, augmented in various ways and prepared for training and testing. Qualitative assessment shows that the model produces detailed and spatially consistent height maps across different urban morphologies with their unique complexities. Quantitative evaluation further confirms the model's effectiveness, with RMSE values of 3.43 m and 3.29 m for Utrecht and Potsdam, respectively, and accuracy rates (δ_1) above 0.43 and 0.50. The results illustrate the feasibility of incorporating KAN layers into encoder–decoder architectures for monocular height estimation. This study highlights KANU-Net as a promising direction for further research in single-image 3D urban reconstruction.

1. Introduction

Height estimation from single optical images has become an important research topic in photogrammetry, remote sensing, and computer vision. Unlike stereo-based or LiDAR approaches, monocular methods aim to reconstruct the vertical dimension of urban environments using only a single image, making them valuable in cases where auxiliary elevation data are unavailable or costly (Amini Amirkolaee & Arefi, 2019). Such capability supports applications including 3D city modeling, urban monitoring, and disaster management (Amirkolaee & Arefi, 2019b). Recent advances in deep learning, particularly convolutional neural networks (CNNs) with encoder–decoder designs, have greatly improved monocular height and depth estimation (Amini Amirkolaee & Arefi, 2021; Amirkolaee & Arefi, 2019a). Owing to their strong feature extraction and generalization capabilities, CNN-based models have also demonstrated promising performance across various remote sensing applications (Ghahrloo & Mokhtarzade, 2025). These models learn spatial and contextual cues from high-resolution aerial and satellite imagery, yet their performance can degrade in heterogeneous urban areas due to limited receptive fields, vanishing gradients, and difficulties in representing complex nonlinear relationships among urban features (Ren, 2024). Consequently, developing alternative architectures capable of capturing richer spatial dependencies remains an open challenge. To address this, emerging studies have proposed Kolmogorov–Arnold Networks (KANs), inspired by the Kolmogorov–Arnold representation theorem. Unlike conventional convolutional or linear layers, KAN layers employ functional basis expansions to model nonlinear interactions, offering potentially higher expressiveness and interpretability (Kundu et al., 2024; Liu et al., 2024). While their application in remote sensing tasks is still limited, early evidence suggests that they can complement CNN-based approaches in spatial prediction problems. In this study, we propose KANU-Net, a U-Net-based architecture enhanced with KAN layers, for height estimation from single RGB images. The model is evaluated on optical imagery from different urban environments to examine its generalization capability. By combining multi-scale feature aggregation with expressive basis function expansions, KANU-Net aims to produce sharper and more consistent height predictions in complex urban settings. The remainder of this paper is organized as follows: Section 2 declares data and study area; Section 3 presents the proposed architecture and methodology; Section 4 contains the experimental outcomes; Section 5 discusses general impressions; and Section 6 concludes the study and outlines future directions.

2. Data and Study Area

Two urban areas were selected: Utrecht, the Netherlands, and Potsdam, Germany. For Utrecht, RGB imagery was acquired from Google Satellite at a spatial resolution of about 8cm, while Potsdam data was obtained from the ISPRS 2D Semantic Labeling Benchmark dataset, also with a 5 cm resolution. Height maps were aligned with the imagery for both regions. According to the different resolutions of the RGB and elevation data, both of them were downsampled to 1m resolution to have the corresponding elevation value of each pixel in the image. The datasets were divided into patches of 256 × 256 pixels for training and evaluation. Using two distinct cities allowed assessment of the model's generalization across different architectural and geographic contexts. Meanwhile, the 256-dimensionality of the dataset patches ensures that there is meaningful data in each patch.

* Corresponding author

https://doi.org/10.5194/isprs-annals-X-4-W8-2025-809-2026

3. Methodology

The proposed methodology for estimating height from a single image can be explained in the following steps: (1) pre-processing, (2) dataset preparation and augmentation, (3) applying the neural network, (4) post-processing.

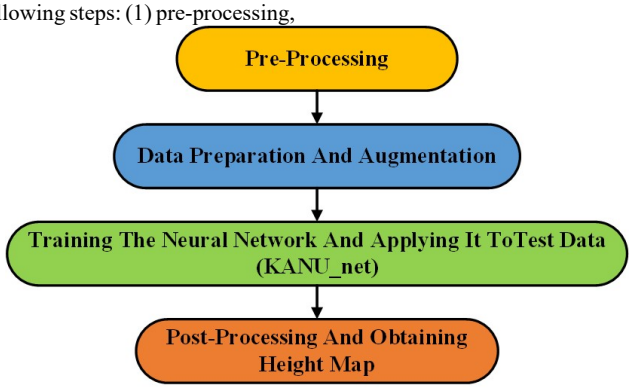

Figure 1. Main steps of the proposed methodology.

3.1 Pre-Processing

To have a uniform and balanced image in the datasets, the pixels with wrong values were identified and properly calibrated. Also, a simple filter was applied to the image. To generate a normal DSM of the area, the statistical median of the lowest elevations was subtracted from the elevation values.

3.2 Data Preparation and Augmentation

Since we are working in an urban area, the minimum size of entities to be detected on the ground was decided to be 1m, which is an appropriate resolution in similar tasks. Furthermore, the dimensions of patches were chosen to be 256*256 to have enough meaning in each patch to make the learning process easier for the network. Data augmentation was applied to increase the volume of the data. Rotation, overlap, flip and random seed points were the augmentation techniques that were applied to the data.

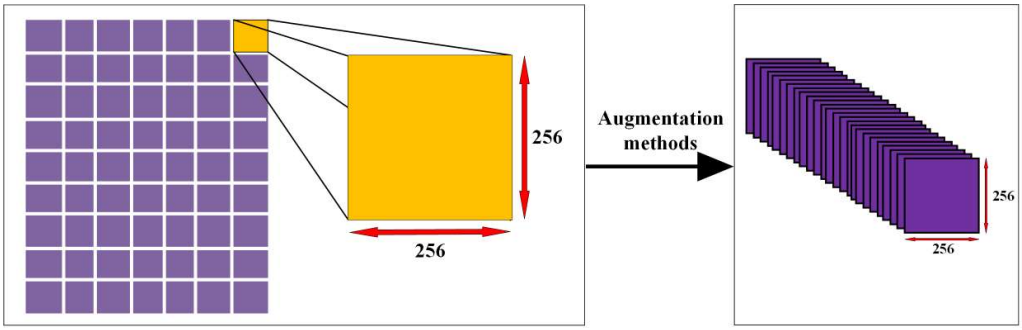

Figure 2. The process of segmenting the RGB image and the corresponding height map into patches with dimensions of 256*256 pixels and generating the complete dataset for training and testing the model, using different augmentation methods.

3.3 The Proposed Neural Network Architecture

To estimate the height of urban features from geospatial imagery, we propose a novel encoder-decoder architecture named KANU-Net. This model is a customized variant of the classical U-Net, in which standard convolutional operations are replaced by FastKANConv layers. These layers leverage basis function expansions based on the Kolmogorov–Arnold Network (KAN) paradigm to enhance representational power and generalization. Unlike traditional neural networks such as Multilayer Perceptrons (MLPs) or Convolutional Neural Networks (CNNs), where the nonlinear activation functions (e.g., ReLU, GELU, or Sigmoid) are fixed and applied uniformly after each linear transformation, the Kolmogorov–Arnold Network (KAN) introduces a fundamentally different approach. In KAN, the nonlinear mappings between neurons are not predetermined but learnable.

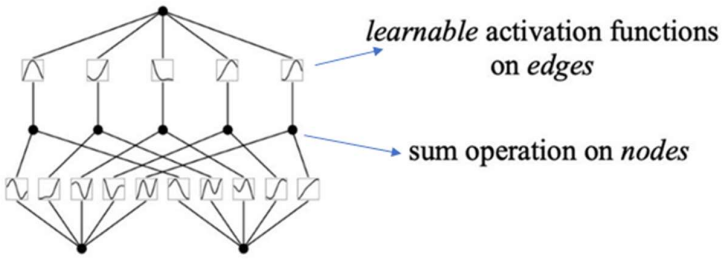

Figure 3. An illustration of the KAN layer structure, where learnable activation functions are applied on the edges and summation operations are performed on the nodes (Liu et al., 2024).

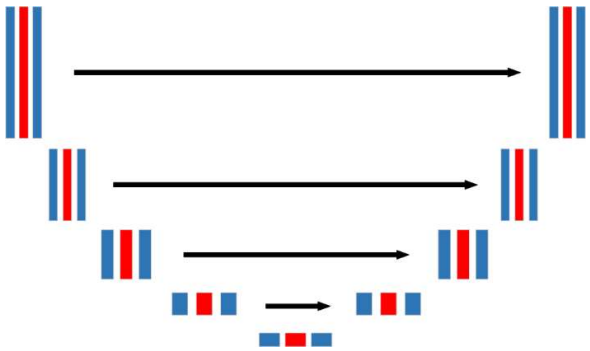

Figure 4. Schematic illustration of the U-Net / encoder–decoder structure with skip connections.

Each FastKANConv layer expands its input features over a set of basis functions—such as Radial Basis Functions (RBFs), Fourier, Polynomial, or B-Spline bases—and combines them through learnable weights before convolution. In this implementation, RBFs are adopted due to their smooth and localized nature, which is particularly effective for representing continuous variables like height or depth. The FastKANConv operation therefore enhances both the expressive power and the generalization ability of the model, capturing complex spatial relationships that standard convolutions might overlook.

Layer	Output Size	Function
INPUT	256×256×3	Input RGB image
ENC1	256×256×64	DoubleConv (FastKANConv ×2)
POOL1	128×128×64	MaxPool2D
ENC2	128×128×128	DoubleConv (FastKANConv ×2)
POOL2	64×64×128	MaxPool2D
ENC3	64×64×256	DoubleConv (FastKANConv ×2)
POOL3	32×32×256	MaxPool2D
ENC4	32×32×512	DoubleConv (FastKANConv ×2)
POOL4	16×16×512	MaxPool2D
BOTTLENECK	16×16×1024	DoubleConv (FastKANConv ×2)
UP1	32×32×512	Bilinear Upsampling
CONCAT1	32×32×1024	Skip Connection with ENC4
DEC1	32×32×256	DoubleConv (FastKANConv ×2)
UP2	64×64×256	Bilinear Upsampling
CONCAT2	64×64×512	Skip Connection with ENC3
DEC2	64×64×128	DoubleConv (FastKANConv ×2)
UP3	128×128×128	Bilinear Upsampling
CONCAT3	128×128×256	Skip Connection with ENC2
DEC3	128×128×64	DoubleConv (FastKANConv ×2)
UP4	256×256×64	Bilinear Upsampling
CONCAT4	256×256×128	Skip Connection with ENC1
DEC4	256×256×64	DoubleConv (FastKANConv ×2)
OUTPUT	256×256×1	FastKANConv (1×1 convolution)

Table 1. Architecture of KANU_Net

The overall architecture of KANU-Net follows the symmetric U-shaped encoder–decoder design. The encoder path progressively downsamples the spatial resolution while increasing the number of feature channels. It consists of five stages, each composed of two sequential FastKANConv blocks followed by a max-pooling operation. These layers extract hierarchical feature representations of urban scenes, from low-level textures to high-level structural patterns. The decoder path mirrors the encoder but performs the inverse process. Each decoding stage begins with bilinear upsampling of the lower-resolution feature map, followed by concatenation with the corresponding encoder features through skip connections. This fusion allows the decoder to recover fine spatial details lost during downsampling. The concatenated features are then refined through two additional FastKANConv blocks, ensuring that both local and contextual information contribute to the reconstructed height map. The final output stage employs a 1×1 FastKANConv layer that maps the last decoder feature map into a single-channel output representing the estimated height. A base update branch is also integrated within each FastKANConv block to stabilize training and retain residual information from the original input features.

KANU-Net maintains the strong spatial inductive bias of convolutional architectures while significantly improving the non-linear representational capacity of the model through the KAN formulation. By learning flexible, data-dependent nonlinear mappings instead of using fixed activations, KANU-Net demonstrates improved robustness and accuracy when generalizing across diverse urban environments. This makes it particularly suitable for height estimation tasks where the visual and structural characteristics of urban features vary significantly across regions and imaging conditions.

3.4 Post-Processing

After obtaining the results, the median value, which was initially subtracted from the height values due to data normalization, was added to the resulting values.

4. Evaluation

The qualitative and quantitative results of applying the trained models to the test area of the two cities under study are presented, along with an analysis of the results.

4.1 Qualitative Result

This part includes the analysis and visual evaluation of the resulting elevation maps. In the following, these maps are presented for the two cities of Utrecht and Potsdam, respectively.

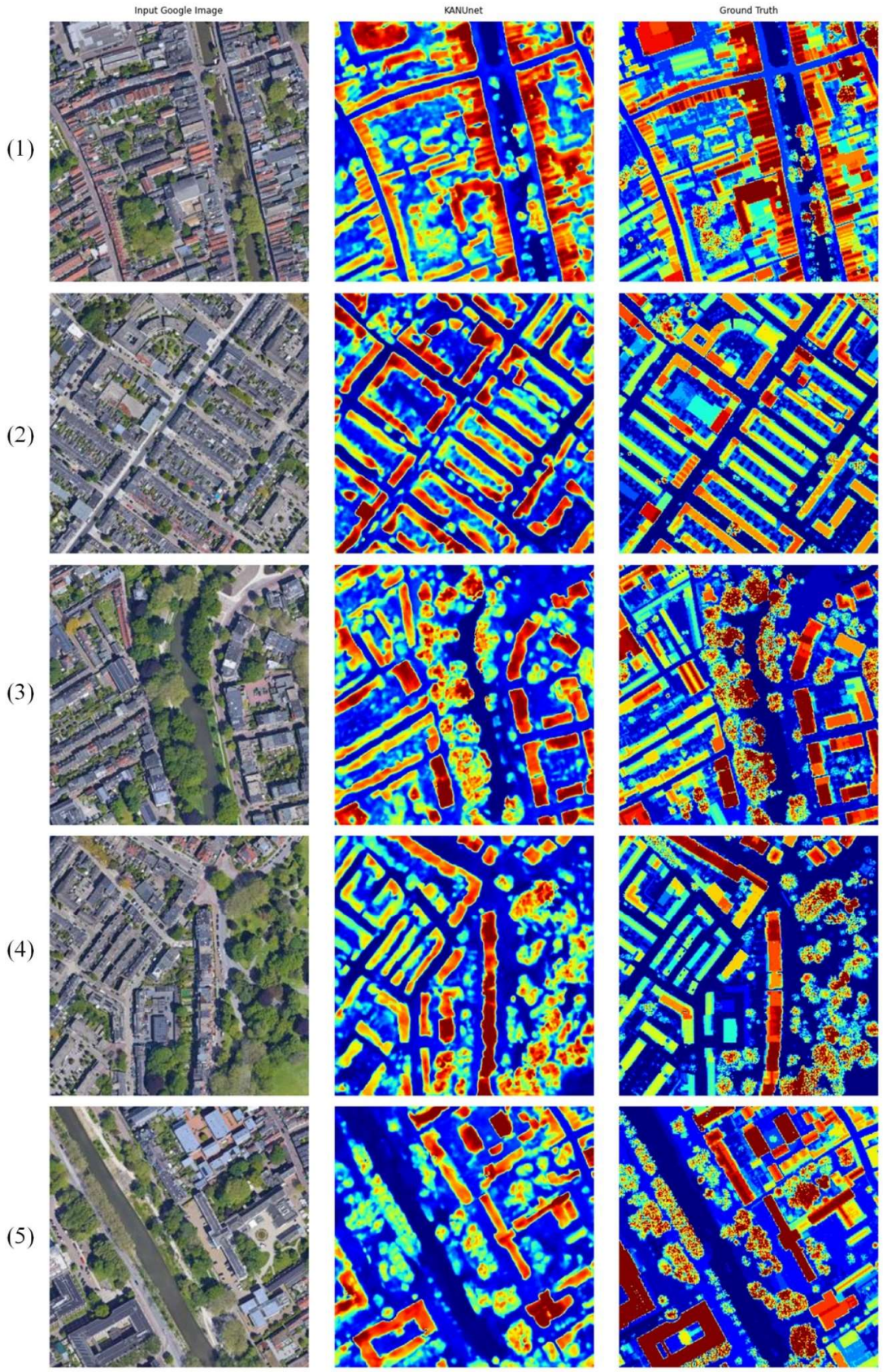

Figure 5. Qualitative results of the KANU-Net model on the Utrecht dataset. The columns show the input Google images, the predicted height maps, and the ground truth height maps, respectively.

In Figure 5 (1), it is demonstrated that the network has been successful in height estimation in building areas. Some ancillary buildings with small dimensions, especially those with lower heights, are ignored by the network. It should also be noted that buildings with roof colors that are different from the asphalt and have more contrast are better extracted. As another point in this figure, dimensional mismatch is observed in some building structures with more shape complexity; the boundary of small and dense blocks is not well identified. However, the results are generally acceptable and appropriate.

The area in Figure 5 (2) includes buildings of various sizes and heights. Higher and bigger ones are well identified, while small components are not extracted properly. In the upper part of the area, a less frequent structure with complexity in shape is observed. The model has performed well in typical parts of these buildings, while in complex sections, there is a discrepancy. Despite the structures being the same color as the ground, the model performed well in extracting buildings in residential areas, especially where there is less vegetation. Also, in some limited cases, the presence of buildings with different geometric shapes and occluded by trees poses challenges in estimating the height for the network in this figure. As shown in the lower part of the image, the overlap of trees with buildings has impeded the accurate extraction of building footprints. However, these challenges do not compromise the validity of the overall results, and they are still acceptable, and of course, the network performance in the total area is good.

Figure 5 (3) and (4), declare that linear and block-shaped buildings in the input image were accurately identified in the model's height estimation map, with their overall shapes well preserved. Narrow pathways (such as streets) in the model output were mostly represented with low or near-zero heights, which is consistent with the expected ground truth. Parts of trees or green spaces adjacent to buildings were occasionally overestimated in height, causing some non-building areas to appear similar to small buildings. There is more vegetation in these images. Trees are well extracted in terms of size because they do not overlap with buildings, but they are identified more as elevation spots than as distinct structures. Compared to the ground truth, the model sometimes exhibited discrepancies of a few meters in estimating the exact height of tall buildings (e.g., high-rise towers), while maintaining the overall height distribution pattern. Furthermore, the details and complexity of the building roofs are not completely extracted. One of the impressive factors in this challenge is resolution. In higher resolutions, the roof details are better extracted.

In Figure 5 (5), it is observed that the model was able to preserve the overall structure of the city with basic details (building boundaries, streets, distances). At the intersection of vegetation with structures, a softening is observed in the identification of the building boundary and the estimation of its true shape. For taller buildings, the model estimated the height fairly well, but slightly under- or overestimated (error of a few meters).

Totally, Figure 5 shows the model performed well in separating and distinguishing the ground surface from buildings, to the extent that most paths, roads, and streets were extracted well. Another noticeable point in the Utrecht results is the effect of shadow on estimated heights. Due to the angle of the sunshine, shadows from various buildings and features are observed in the Utrecht images, which affects the color values of the pixels. However, in the elevation outputs, we see good results and changing the pixel values from shadows did not affect the recognition and performance of the network. So, it represents one of the strengths of the proposed network.

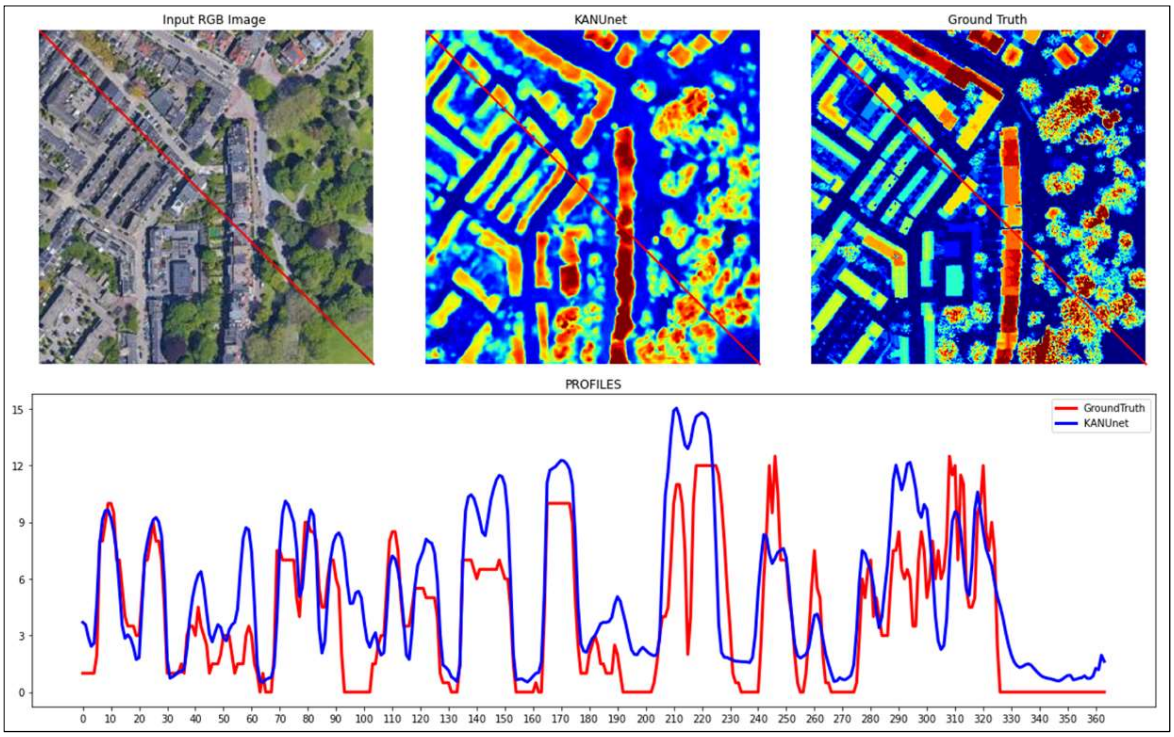

Figure 6. Profile comparison of height estimation results for a sample area in Utrecht. The top row shows the input RGB image, the height map predicted by the proposed KANU-Net model, and the corresponding ground truth. The bottom plot illustrates the elevation profiles extracted along the diagonal line, demonstrating the consistency of the predicted KANU-Net heights (blue profile) with the reference data (red profile).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W8-2025-809-2026 | © Author(s) 2026. CC BY 4.0 License.

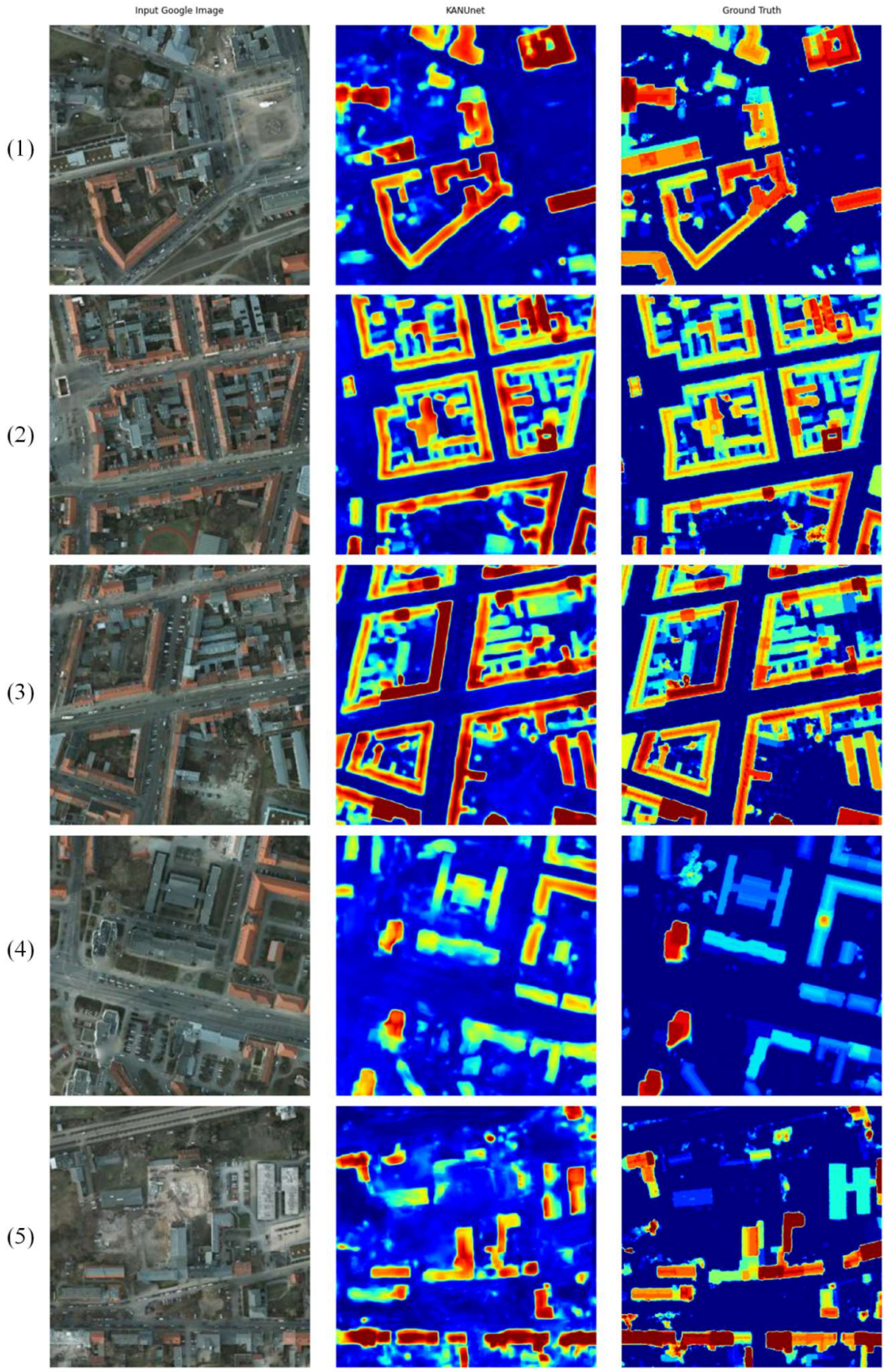

Figure 7. Qualitative results of the KANU-Net model on the Potsdam dataset. The columns show the input RGB images, the predicted height maps, and the ground truth height maps, respectively.

In Figure 7 (1), the model worked well on large buildings and preserved the original height structure, with good accuracy on small buildings and boundaries. Separation of ground and non-ground has been done properly, and the total shape of the buildings is extracted well. Although in the middle of the image, the buildings are well identified by the good contrast in the color of their roofs, in the left and lower corner of the image, some buildings are not properly identified due to the similarity of their color to the asphalt, and their boundaries are estimated to be at the same height as the ground.

Figure 7 (2) contains the main buildings with auxiliary structures in the middle of them. The main buildings' shape and boundaries are extracted well, and despite the 1-meter resolution, we can see that the auxiliary buildings, such as skylights, interior courtyards, or irregular roofs, have reasonable results. Maybe the model sometimes had difficulty fully separating these parts, but the general elevation results are acceptable in this resolution. Additionally, the roof details are being relatively identified, and it is because of the color contrast and the omission of the tilt and relief displacement impact from the total RGB image of this city. Another considerable point is the well extraction of the ways and streets in this figure.

In Figure 7 (3) on the estimated height map, plus the good street and building detail extraction, we can see cars on the street. High contrast of these objects is the reason for their extraction. The model could distinguish the cars and the roads due to their color. Of course, the absence of vegetation and any other impediments helped with the car identification too.

Due to the small dimensions and low height of secondary buildings, they are expected to be omitted or ignored by the network, but in both Figure 7 (2) and (3), low-rise secondary buildings have been detected, with their heights estimated properly by the model. Although slight boundary smoothing is observed around these structures.

In Figure 7 (4), the RGB image contains different features, including buildings in various shapes, heights, sizes and roof color contrast, low vegetation (e.g., grass) and vehicles. The model performance in estimating buildings with high contrast is good, and in estimating buildings with lower contrast is acceptable. Areas covered with low-height vegetation, despite having a different color from the bare ground, have been estimated with ground-level heights, which is actually true. Even vehicles have been partially identified in the output. However, the main challenge in this image is the presence of very tall buildings, whose heights have been estimated with a deviation of several meters from the ground truth.

Results in Figure 7 (5) show that despite the low color contrast between ground and structures, the network performed successfully in recognition and elevating the features in the area. There are some mismatches in very low contrast buildings, but the total performance is acceptable. Also, vegetation couldn't confuse the model in identifying the terrain and its elevation, and it is a positive point.

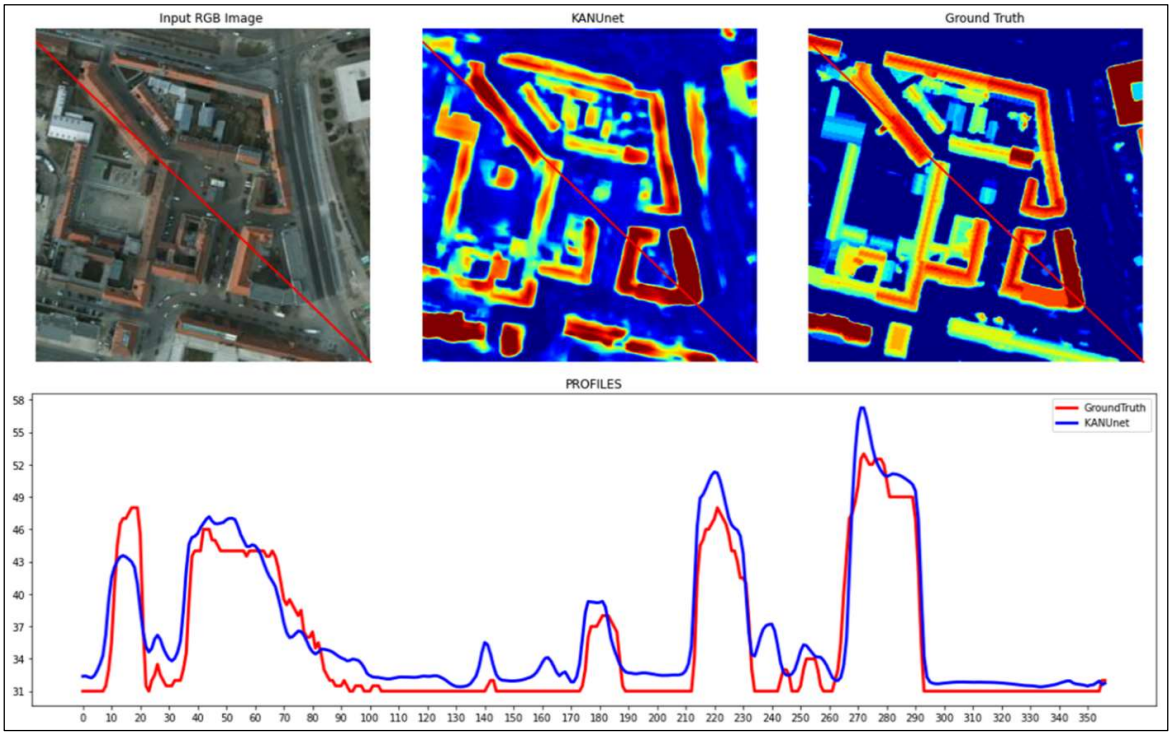

Figure 8. Profile comparison of height estimation results for a sample area in Potsdam. The top row shows the input RGB image, the height map predicted by the proposed KANU-Net model, and the corresponding ground truth. The bottom plot illustrates the elevation profiles extracted along the diagonal line, demonstrating the consistency of the predicted KANU-Net heights (blue profile) with the reference data (red profile).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W8-2025-809-2026 | © Author(s) 2026. CC BY 4.0 License.

4.2 Quantitative Results

To quantitatively evaluate the performance of the proposed KANU-Net, we computed standard error metrics, including the Root Mean Square Error (RMSE), on the test subsets of the Utrecht and Potsdam datasets. Table 2 summarizes the results.

$$RMSE = \sqrt{\frac{1}{n}\sum(H_r - H_e)^2} \quad (1)$$

$$\delta_i = max\left(\frac{H_r}{H_e}, \frac{H_e}{H_r}\right) < 1.25^i \ , i \in \{1,2,3\} \quad (2)$$

Where n = number of pixels
Hr = reference height pixel value
He = estimated height pixel value

In Eq. (1), which is related to the difference between reference and estimated pixel value, the lower the RMSE value, the higher the accuracy. In Eq. (2), which is based on the ratio of the reference and estimated pixel values, the higher the δ value, the better the result.

Metrics / Dataset	RMSE (m)	δ_1	δ_2	δ_3
Utrecht	3.4281	0.4319	0.6603	0.7930
Potsdam	3.2850	0.5051	0.6951	0.7872

Table 2. Quantitative metrics of depth estimation

5. Discussion

In comparing the characteristics of the two datasets and the outcomes obtained for the two cities, the following observations can be made: (1) The structure of these two cities is different from each other. The majority of the buildings in Utrecht are of modest dimensions, higher elevation and low color contrast. Most of the roofs of the buildings are not very complex, but sometimes there are some complexities in the shape of the buildings. In addition to the existing structural density, the color image prepared for this city was taken during the leafy season of the trees, and in the results presented, the effect of this dense vegetation cover was discussed in detail. In the Potsdam city data, however, the urban density is lower than that of Utrecht. The buildings are more isolated and larger, and the roofs of the buildings have higher contrast and better shape differentiation. There is also no dense vegetation in this city data, and only limited vegetation with very low height is visible, which leads to a reduction in errors caused by the overlap of vegetation with other features. (2) Another difference between these datasets is the presence of tilt and relief displacement in the color image. This impact has been removed from the Potsdam images, while it exists in the Utrecht dataset images. The absence of tilt and relief displacement and the overall orthophoto nature of the color image of Potsdam city maintain accuracy in identifying the correct location of the feature boundary and bring the shape and area of buildings closer to their reality. Despite these challenges, the network trained well with both datasets and produced acceptable and appropriate results from these cities. It means that the network can operate in different areas with different characteristics.

6. Conclusion

In this work, we investigated the use of KANU-Net, a U-Net variant enhanced with Kolmogorov–Arnold Network layers, for monocular height estimation from single optical images. The model was applied to two urban datasets, Utrecht and Potsdam, with 1 m resolution imagery, allowing us to examine its behavior across different urban morphologies. Visual evaluation demonstrated that KANU-Net can produce detailed and spatially consistent height maps, capturing structural variations in dense urban areas as well as more regular building patterns. These findings confirm the feasibility of integrating KAN-based operators into encoder–decoder networks for height estimation tasks. While the current study was limited to two datasets, the results provide a promising step toward the use of functional basis expansions in monocular 3D reconstruction. Future work will focus on expanding the evaluation to larger and more diverse datasets and incorporating quantitative comparisons with baseline architectures.

7. Acknowledgement

The authors would like to thank JaouadT for sharing the KANU-Net implementation on GitHub.

8. References

Amini Amirkolaee, H., Arefi, H., 2019. 3D change detection in urban areas based on DCNN using a single image. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *42*, 89–95.

Amini Amirkolaee, H., Arefi, H., 2021. Generating a highly detailed DSM from a single high-resolution satellite image and an SRTM elevation model. *Remote Sensing Letters*, *12*(4), 335–344.

Amirkolaee, H. A., Arefi, H., 2019a. Convolutional neural network architecture for digital surface model estimation from single remote sensing image. *Journal of Applied Remote Sensing*, *13*(1), 016522–016522.

Amirkolaee, H. A., Arefi, H., 2019b. Height estimation from single aerial images using a deep convolutional encoder-decoder network. *ISPRS journal of photogrammetry and remote sensing*, *149*, 50–66.

Ghahrloo, M., Mokhtarzade, M., 2025. Earthquake-induced building damage detection using the fusion of optical and radar data in intelligent systems. *Earth Science Informatics*, *18*(1), 62.

Kundu, A., Sarkar, A., Sadhu, A., 2024. Kanqas: Kolmogorov-arnold network for quantum architecture search. *EPJ Quantum Technology*, *11*(1), 76.

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., Tegmark, M., 2024. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*.

Ren, C., 2024. Eite-Mono: an extreme lightweight architecture for self-supervised monocular depth estimation. *IEEE Access*.