This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3564820

Depth2Elevation: Scale Modulation with Depth Anything Model for Single-view Remote Sensing Image Height Estimation

Zhongcheng Hong, Tong Wu, Zhiyuan Xu, and Wufan Zhao

***Abstract*—Accurate terrain elevation estimation from remote sensing data is essential for a multitude of geographic applications. Specifically, image-based elevation estimation has garnered growing attention due to advancements in optical sensor development and automated analysis algorithms, such as machine learning. In this context, deep learning methods, including Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have recently enhanced the feature extraction ability and estimation accuracy of this task. Despite the distinct advantages afforded by each architectural paradigm, current methods are frequently impeded in their ability to discern subtle height variations within complex scenes and are ill-equipped to effectively tackle the extraction of features across both large and small scales. Although vision foundation models have shown significant advances in remote sensing analysis, their effectiveness for height estimation remains unexplored. In this study, we introduce the foundation model in the field of elevation estimation and propose a novel Depth to Elevation (*Depth2Elevation*) model, marking the first application of the Depth Anything Model (DAM) to height estimation in remote sensing images. First, we introduce the scale modulator for modulating partial encoders in the original DAM, which enables DAM to capture subtle representations of localized objects at different scales. Secondly, we further enhance the model's representational capability by using a resolution-agnostic decoder architecture, which enables DAM to learn features at different spatial scales efficiently. We conducted comprehensive experiments on several benchmark datasets. Compared to strong baselines, our method achieves an average relative improvement of at most 42% on the latest large-scale benchmark dataset GAMUS and shows the best generalization ability across different scenarios.**

***Index Terms*—single-view image height estimation, depth anything model (DAM), vision transformer, scale modulation**

I. Introduction

In today's era of rapid development of information technology, the accurate measurement and analysis of surface features are essential for urban planning, environmental monitoring, disaster assessment, and geographic information systems (GIS) [1, 2]. Among these tasks, terrain elevation estimation (also named height estimation) plays a crucial role in understanding the spatial distribution of surface features and supporting various scientific and engineering applications. Moreover, recent studies have shown that digital surface model (DSM) is useful for more challenging tasks such as semantic labeling and change detection [3, 4, 5]. Estimation of terrain elevation involves generating DSM corresponding to a given image. Traditional elevation estimation techniques, including stereo image pairs, structure from motion (SfM), and light detection and ranging (LiDAR) laser scanning [6, 7], achieve high accuracy but are frequently constrained by substantial time and cost requirements, limiting their practicality for large-scale applications.

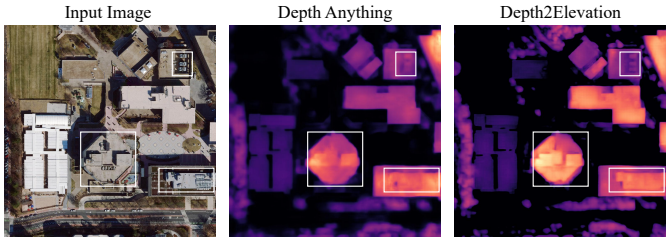

Fig. 1. **Comparison of the prediction results of DAM directly fine-tuned and our proposed Depth2Elevation method.** Our proposed Depth2Elevation has more accurate boundaries and local heights in the height prediction of the roof structure.

With advancements in remote sensing technology, particularly the widespread availability of high-resolution satellite and aerial imagery, image-based elevation estimation methods have garnered significant attention [8, 9, 10, 11, 12]. These approaches predict surface elevation information by leveraging features extracted from 2D images. However, accurately estimating elevation from a single-view image presents considerable challenges due to factors such as viewing angles, occlusions, lighting conditions, and the complexity of surface characteristics. The emergence of deep learning in recent years has introduced innovative solutions to these challenges, with numerous approaches employing CNNs [13, 14, 15, 16] and ViTs [17, 18, 19, 20, 21]. The former approach extracts local features from an image through stacked convolutional operations, while the latter captures global features by modeling long-range dependencies. Despite the unique advantages of each model, their capacity to capture subtle changes in localized building heights in high-resolution remote sensing images is limited, which restricts their application to diverse remote sensing datasets. Furthermore, unlike natural images, high-resolution remote sensing images typically contain com-

Zhongcheng Hong, Zhiyuan Xu, and Wufan Zhao are with the Urban Governance and Design Thrust, Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, China (e-mail: Zhongchengh@hkust-gz.edu.cn, wufanzhao@hkust-gz.edu.cn). Zhiyuan Xu is also with the School of Computer Science, University of Bristol, Bristol, BS8 1QU, UK (Email: zhiyuan.xu@bristol.ac.uk). Wufan Zhao is the corresponding author.

Tong Wu is with the School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Private Bag 92006, Auckland 1142, New Zealand (e-mail: tong.wu@autuni.ac.nz).

plex scenes with multiple objects exhibiting significant scale variation. A single model structure struggles to meet the feature extraction demands of both large-scale and small-scale objects effectively.

Substantial progress has been achieved in the domain of computer vision due to the deployment of expansive foundational models, and this progress has catalyzed the development of analogous methodologies within the field of remote sensing [22, 23, 24]. These models are pre-trained on large-scale datasets and fine-tuned on the downstream tasks. Despite its remarkable capabilities, the full potential of the foundation model remains largely untapped in remote sensing image height estimation applications. While the Depth Anything Model (DAM) [25] has demonstrated widespread success in monocular depth estimation tasks, it faces notable limitations in meeting the specific requirements of height estimation. The model's output is confined to relative depth values for image pixels, falling short of providing the absolute height measurements crucial for height estimation tasks. Furthermore, DAM predicts the depth of the entire image, but the height map represented by DSM focuses more on the height of objects above the ground. As shown in Fig. 1, directly fine-tuning DAM (freezing encoder and only training prediction head) to predict heights results in blurred boundaries of some objects and overly smooth local heights of the objects themselves. This is fatal for the height estimation task because the object itself (especially buildings) requires clear local height to obtain its shape and structure.

To address the aforementioned challenges, we introduce DAM to the task of remote sensing image height estimation and propose a comprehensive framework that encompasses three key innovations. First, we incorporate a scale modulator into the original DAM to modulate a portion of the encoder, enabling the model to capture subtle representations of local objects at different scales, thereby providing feature information at different scale levels for the subsequent resolution-agnostic decoder. Second, we further enhance the model's representational capabilities by employing a resolution-agnostic decoder architecture, allowing the model to effectively handle features at different scales, thus dealing with images of varying spatial resolutions. Additionally, during the training process, we design a multiple loss which combines mean square error (MSE) loss, gradient loss (GradLoss) and scale-invariant loss (SiLoss), enabling the model to preserve geometric edge structures, significantly distinguish between the ground and objects on the ground, and mitigate the impact of extreme height values or outliers on training. The contributions of this research are as follows:

- We design a novel height estimation network based on DAM, which is the first application of the vision foundation model in the field of remote sensing image height estimation.
- We propose a novel scale modulator that enables the DAM to accurately capture the subtle representations in localized object heights at different scales so as to provide feature information of the corresponding scale for the subsequent resolution-agnostic decoder.
- We propose a resolution-agnostic decoder architecture that outputs predicted height maps at different scales and calculates the loss separately. This strategy enhances DAM's ability to characterize features at different scales, making it more robust in diverse spatial resolution remote sensing scenarios.
- We design a multiple loss that combines MSE loss, GradLoss, and SiLoss. This approach enables the model to enhance the capture of object geometric structures and reduces the interference caused by extreme height values and outliers on the model's learning of normal height values.

II. Related Work

A. Monocular Depth Estimation

Monocular depth estimation (MDE) is a fundamental problem in computer vision, with broad applications in robotics, autonomous driving, virtual reality, etc. Much work has focused on improving the network structure. In the work of Eigen et al. [26], they first used a convolutional neural network consisting of two modules for depth map prediction, using dense depth maps obtained from depth cameras as supervision. Furthermore, since ResNet has demonstrated excellent capabilities as an encoder in computer vision tasks, Laina et al. [27] replaced the encoder, leveraging ResNet to enhance the network's ability to extract features and learn the mapping relationship between depth maps and input images. Lee et al. [28] optimized the decoder, and unlike the standard upsampling layer, they used a locally planar guidance layer to guide the upsampling of feature maps to full resolution. Additionally, since the depth range of objects distributed in different images varies, Bhat et al. [29] proposed a Transformer-based structure, dividing the depth range into many intervals and adaptively estimating the depth median for each image. Compared with optimizing the neural network structure, Nddepth [30] introduced a novel physics (geometry)-driven deep learning framework for monocular depth estimation, assuming that 3D scenes are constituted by piece-wise planes. And Gedepth [31] proposed a ground embedding module to decouple camera parameters from pictorial cues, enhancing the generalization capability of monocular depth estimation.

Recently, with the prevalence of vision foundation models, there have been related studies on monocular depth estimation [25, 32]. Depth Anything V1 [25] enhanced the performance of monocular depth estimation models by leveraging large-scale unlabeled data, emphasizing the importance of untagged real data in bridging domain gaps and enhancing scene coverage. Depth Anything V2 [32] made improvements based on V1, replacing all real labeled image data with synthetic images, enlarging the size of the teacher model, and training the student model with large-scale pseudo-labeled real images, thereby significantly improving details and robustness, and achieving faster inference speed and fewer parameters. However, as far as single-view height estimation is concerned, there is no method involving vision foundation models.

B. Single-view Height Estimation

Deep learning has catalyzed remarkable advancements in the realm of single-view height estimation, with convolu-

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3564820

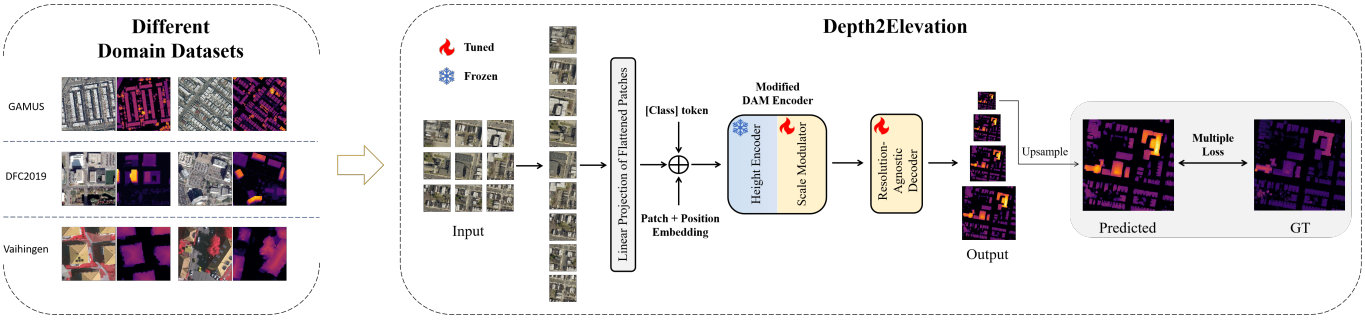

Fig. 2. **Architecture of Depth2Elevation.** Our Depth2Elevation is capable of adapting to data from different domains, which means that we have tried training with datasets from various domains, and it has achieved impressive results across all of them. We modify DAM's encoder as our height encoder. The input image is processed into multiple patches. Then the height encoder and scale modulator extracts features from the input patches and the resolution-agnostic decoder converts these features into the predicted height map. Subsequently, multiple loss is used for supervision on predicted height maps of different scales.

tional neural networks (CNNs) emerging as a cornerstone for feature extraction and end-to-end height estimation from a solitary image [9, 13, 13, 14, 33, 34, 35]. Pioneering work IM2HEIGHT model by Mou and Zhu [33] has set a precedent, using a fully convolutional network architecture to achieve mapping from a single-view image to a digital surface model (DSM). Since then, a number of CNN-based methods have emerged. Karatsiolis et al. [34] proposed the IMG2nDSM model, which combined a deep learning architecture to estimate the height of buildings and vegetation from a single aerial RGB image, demonstrating good performance. In addition, the IM2ELEVATION model [35] used single-view images for height estimation through multi-sensor fusion, demonstrating good performance. Li et al. [13] proposed a solution based on a deep ordinal regression network and introduced the ASPP module into the network to realize multi-scale feature extraction. Furthermore, Gated Feature Aggregation method proposed by Xing et al. [14] enhanced the ability of single-view height estimation by effectively combining low-level and high-level features, especially in preserving object boundaries and contours. LUMNet model [9] significantly improved the accuracy of height estimation by combining land use knowledge and multi-scale feature extraction.

As Transformer emerges in the field of computer vision, some methods have proposed Transformer-based networks [17, 18, 19, 20]. Chen et al. [17] proposed HTC-DC Net, which consists of a feature extractor, an HTC-AdaBins module, and a hybrid regression process. It aimed to solve the long-tail distribution problem in monocular height estimation and determine the bins adapted to each input image through the classification stage. Xiong et al. [18] proposed a scale-deformable convolution module to enhance the scale change problem in Transformer-based height estimation tasks. In addition, Wu et al. [19] proposed a new Transformer-based architecture called HeightFormer, which includes a multi-scale visual Transformer as an encoder and decoder, and a bilateral feature pyramid fusion scheme to enhance global and local information reconstruction. Chen et al. [20] proposed a network that combines multi-level interactions and image adaptive classification-regression. The network improves the quality of instance-level height estimation through a multi-level interaction background and image adaptive classification-regression height generator, which significantly improves edge sharpness.

Some methods attempt unsupervised learning. Zhao et al. [15] proposed a semantically-aware unsupervised domain adaptation method, which includes 2 stages: image translation and multi-task representation learning, aiming to improve the performance of height estimation from single-view orthophotos under unsupervised domain adaptation and alleviate the training limitations of nDSM data. In addition, Zhao et al. [12] also proposed a multi-scale refinement network based on contrastive learning for estimating height from a single aerial image. This method uses a gradient-based self-supervised learning network and momentum contrastive loss to extract geometric information from unlabeled images in the pre-training stage, and uses a local implicit constraint layer in the supervised network to refine the high-resolution features of height estimation.

Although the above-mentioned methods have made remarkable achievements in feature extraction and end-to-end height estimation from a single image, they are still limited in their ability to capture subtle height changes in complex scenes, which restricts their application in diverse spatial resolution remote sensing datasets and is a major challenge for the task of height estimation from remote sensing images.

III. Methodology

A. Modal Overview

Fig. 2 shows the overall framework of the proposed Depth2Elevation model. Specifically, we designed a scale modulator and a resolution-agnostic decoder to transfer the depth estimation capability of DAM to the height estimation of a single-view image. The model is adaptable to datasets of different scales and takes a single-view image as input, extracting features through the height encoder and scale modulator. Then the features perceived by scale modulator are fed into the resolution-agnostic decoder. Finally, the resolution-agnostic decoder decodes the features and outputs the height map.

B. Scale Modulator

The scale modulator represents a refinement of the DAM encoder. As shown in Fig. 2, in the Depth2Elevation model,

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3564820

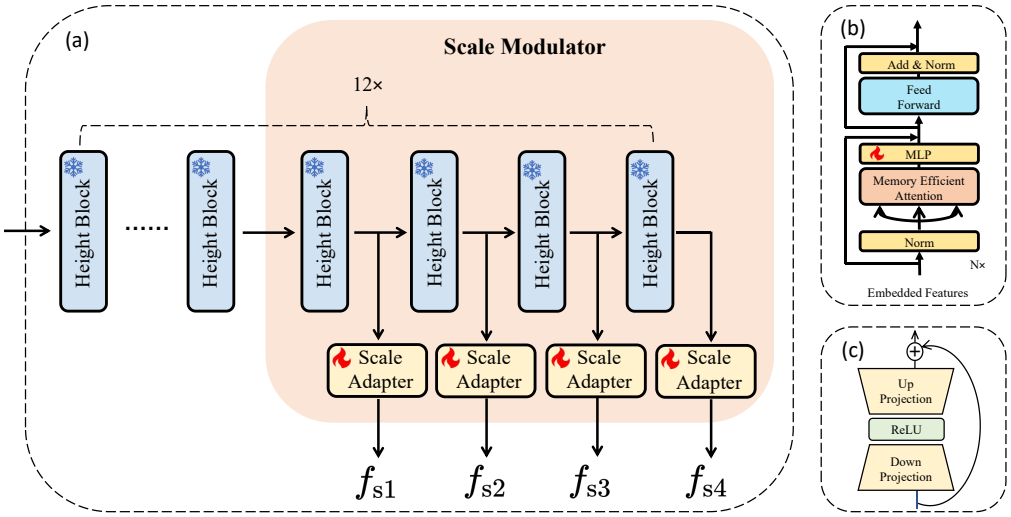

Fig. 3. **Details of Depth2Elevation encoder and scale modulator.** *(a)*: Architecture of Scale Modulator. The embedded features of the last four blocks of the height encoder are input into the scale adapter to further extract relevant features that adapt to the decoder scale. *(b)*: The architecture of height block. *(c)*: The architecture of scale adapter.

similar to DAM, the height encoder employs ViT-B [36] as its backbone, segmenting the input image into fixed-size patches. These patches are then projected into a high-dimensional space and transformed into embedded features. This process converts the 2D image into a sequence of linearized, high-dimensional patch embeddings, which serve as input for subsequent processing by the transformer layers. The network further augments each patch embedding with positional encoding to maintain the spatial position information of the patch within the image. This step is crucial for the model to comprehend the relative or absolute position of each patch in the original image, thus preserving the spatial structure information. Following this, the height encoder, as shown in 3 (a), consists of a series of twelve height blocks that we modify from transformer block of the DAM encoder. The detailed architecture of each height block is shown in Fig. 3 (b). Compared with the transformer block in the original DAM encoder, we integrate a trainable Multi-Layer Perceptron (MLP) after the attention residual connection of each transformer block to ensure efficient and stable model training.

Next, we develop a scale modulator to bolster the synergy between the encoder's output embedded features and the resolution-agnostic decoder and encourage the model to capture both semantic information and fine-grained geometric details. The architecture of scale modulator is illustrated in Figure 3 (a). After the input has been processed by the height encoder, we feed the embedded features from the last four blocks into the scale adapter (see Fig. 3 (c)). These embedded features are destined for the resolution-agnostic decoder, which will generate multi-scale height maps. The specific architecture of scale adapter is shown in Fig. 3 (c). The up projection and down projection of scale adapter are linear layers. Down projection maps the input embedded feature dimension to 128 dimensions. After passing through the activation function ReLU, up projection maps the compressed embedded feature dimension back to the original dimension and connects it with the input feature. This architecture can reduce the number of parameters and computational complexity. Due to the fewer parameters that need to be trained, the model can adapt to new tasks more quickly. Moreover, through residual connections, the original features can be directly conveyed, preventing information loss. The scale adapter is designed to learn the specific feature information corresponding to the height map of each scale, thereby assisting the resolution-agnostic decoder in acquiring the necessary features at the appropriate scales. This enhancement ensures that the Depth2Elevation model can effectively integrate scale-specific information and accurately capture the subtle changes in localized object heights, leading to more accurate and detailed height estimations across varying scales.

C. Resolution-Agnostic Decoder

The structure of the resolution-agnostic decoder is illustrated in the left column of Fig. 4. It receives embedded features corresponding to scale via the scale modulator. The projection block then maps these embedded features back to a 2D space, followed by convolution operations performed through the refine block. The shallow convolution layer simultaneously processes features from the deep convolution layer and those output by the scale modulator.

Specifically, the structure of the projection block is depicted in the upper right pane of Fig. 4. The embedded features received from the scale modulator are transformed into 2D features after being processed by a linear layer and Gaussian Error Linear Units (GELU) activation, facilitating subsequent convolution operations. The primary function of the projection block is to map embedded features encoded by the encoder from a 1D latent feature space to a 2D space.

The structure of the refine block is shown in the lower right pane of Fig. 4. Here, f_n represents the features of

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3564820

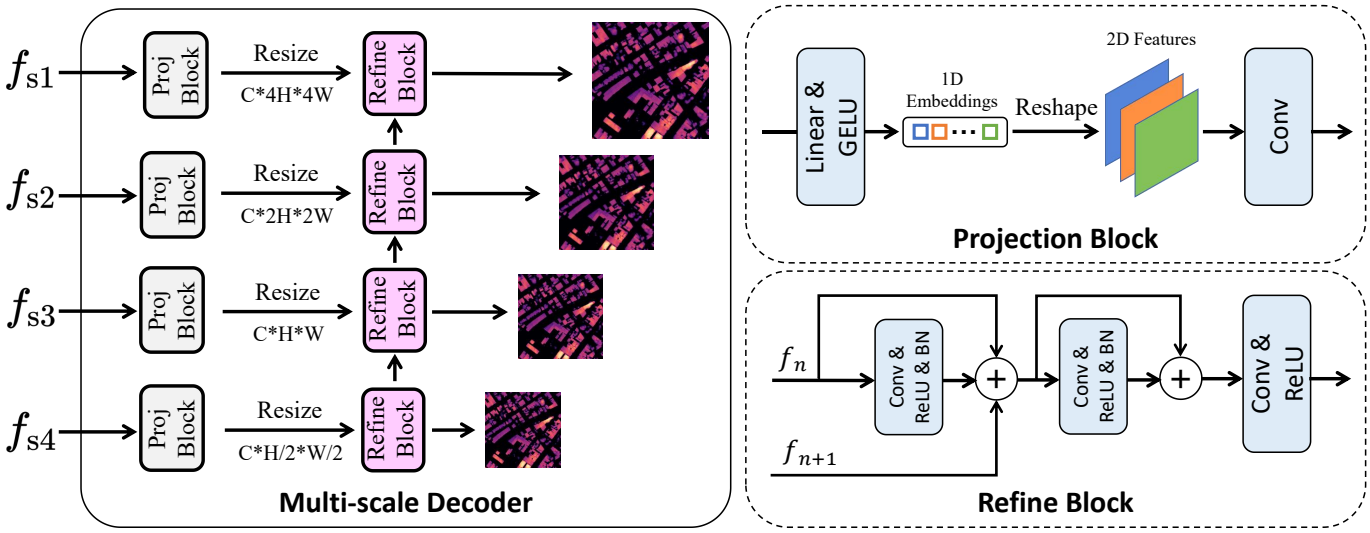

Fig. 4. **Architecture of resolution-agnostic decoder.** The resolution-agnostic decoder receives the features of the multiple branches output by the scale modulator and decodes them into height maps of four scales.

the current layer, while f_{n+1} denotes the features from the deeper layer. With the exception of the deepest layer, each refine block has two input features. The features f_n of the current layer are combined with the deeper features f_{n+1} after a residual connection, enabling the full fusion of features across different scales. This fusion retains high-level semantic information while leveraging low-level positional information, thereby enhancing the model's ability to recognize objects of varying scales in the image. The fused features continue to be input into the residual connection. This residual connection helps alleviate the gradient vanishing problem in deep network training, allowing the network to learn features more deeply while maintaining the stability of training. The final features are passed through the convolution layer and ReLU to output the predicted height map.

During the training phase, the resolution-agnostic decoder outputs height maps at four different scales. Each height map is used to calculate the loss; however, we do not compute the loss directly at the current scale. Instead, we perform a multi-scale alignment loss which upsample the height map to match the input image resolution before calculating the loss. The resolution-agnostic decoder enables the model to capture features across various scales, as a specific height value at a low resolution may correspond to an entire object (such as a building) at a high resolution. This allows the model to learn consistent height cues from objects of different scales. Optimizing at each scale promotes the fusion of features across these scales. Shallow features typically contain more fine-grained geometric details, while deep features provide richer semantic information. By combining these different levels of features, the overall performance of the model can be significantly improved.

D. Multiple Loss

To enhance the performance of the Depth2Elevation model for height estimation tasks, we employ a combination of MSE loss, SiLoss L_{si} [26], and GradLoss L_{grad} [37]. The MSE loss is instrumental in guiding the model to learn accurate height measurements by minimizing the discrepancies between predicted and GT values. The MSE loss we used is as follows:

$$L_{ai} = \frac{1}{N}\sum_{i=1}^{N} \text{MSE}(p_i, t_i) \cdot m \tag{1}$$

where

$$\text{MSE}(p_i, t_i) = \frac{1}{|m|}\sum_{j \in m} (p_i[j] - t_i[j])^2 \tag{2}$$

Among them, p_i is the ith predicted value (after upsample), t_i is the Ground Truth, m is the outlier filtering mask, $|m|$ is the number of valid pixels in the mask, and N is the number of scales.

We also incorporate SiLoss due to the relatively small height values characteristic of most ground-level objects. This loss is computed in logarithmic space, which mitigates the undue influence of extreme height values or outliers on the overall loss. As a result, the model is better equipped to learn the nuances of smaller height values and to align more closely with the natural distribution of heights. The SiLoss formula is as follows:

$$L_{si} = \frac{1}{N}\sum_{i=1}^{N} \lambda_s \left(10 \cdot \sqrt{\text{Var}(g_s) + \beta \cdot (\text{Mean}(g_s))^2}\right) \cdot m \tag{3}$$

where

$$g = \frac{1}{|m|}\sum_{j \in m} [log(p_i + \epsilon) - \log(t_i + \epsilon)] \tag{4}$$

Where p_i is the ith predicted value (after upsample), t_i is the Ground Truth, m is the outlier filtering mask, $|m|$ is the number of valid pixels in the mask, and N is the number of scales. We set ϵ to $1e^{-7}$ to avoid zero in the logarithm, β to 0.15, $\text{Var}(g)$ represents the variance of g, and $\text{Mean}(g)$ is the mean of g.

Furthermore, given the significant height disparities between objects (predominantly buildings) and the ground, we introduce GradLoss into our model. This loss function evaluates the gradient differences for each pixel across both horizontal and vertical axes, enabling the model to more effectively capture geometric details and the precise delineation of object boundaries. The GradLoss is as follows:

$$L_{grad} = \frac{1}{N}\sum_{i=1}^{N}\lambda\left(\frac{1}{|m|}\sum_{i=1}^{m}|G^{s}_{\text{pred}}(i) - G^{s}_{\text{gt}}(i)|\right)\cdot m \quad (5)$$

Here, G^{s}_{pred} and G^{s}_{gt} are the gradients of predicted height map and ground truth at scale N respectively, m is the outlier filtering mask, $|m|$ is the number of valid pixels in the mask, and N is the number of scales. We set λ to $1e^{-3}$.

This multifaceted approach to loss functions ensures that the Depth2Elevation model is robust and adept at handling the complexities of height estimation tasks.

Finally, the overall Loss L is as follows:

$$L = \gamma L_{ai} + \delta L_{si} + \mu L_{grad} \quad (6)$$

Here we set γ, δ, μ to 1, 1, 0.05 respectively.

IV. Experiments

A. Datasets

In this study, the three publicly available height estimation datasets we selected encompass different sensor types, such as the high-resolution satellite imagery of the GAMUS dataset, the aerial imagery of the DFC2019 dataset, and the airborne data of the Vaihingen dataset. These datasets also cover a variety of geographical areas, ranging from urban to rural settings, ensuring the applicability of the model in different environments. The datasets also feature different spatial resolutions and data volumes, which test the model's ability to handle varying scales and data sizes. Such a large-scale dataset offers rich materials for training deep models, ensuring the effectiveness of model transfer learning, and provides strong evidence for the model's practicality and generalization ability. Moreover, the diversity and breadth of these datasets lay a solid foundation for the future expansion of the model to other geographical areas, indicating the potential of the Depth2Elevation model to be applied in a wider range of geographical information acquisition scenarios.

1) Geometry Aware Multi-modal Segmentation (GAMUS) dataset [38]. The GAMUS dataset is a novel benchmark dataset designed for remote sensing data with a resolution of 0.33m, containingcolor images (RGB) and normalized digital surface model (nDSM) data. It contains 8724 tiles of 1024×1024 resolution collected from five different cities. The dataset is divided into a training set (5004 tiles), a validation set (859 tiles), and a test set (2861 tiles).
2) DFC2019 dataset [39]. The DFC2019 dataset is organized by the Image Analysis and Data Fusion Technical Committee (IADF TC) of the IEEE Geoscience and Remote Sensing Society (GRSS), the Johns Hopkins University (JHU), and the Intelligence Advanced Research Projects Activity (IARPA). It consists of 2783 training images and 100 test images. Since the test images are not publicly available, we use 2783 images for experiments. The size of each RGB image and the corresponding nDSM is 1024×1024. We use 2226 images for training and 557 images for testing. We downsample the images to 518×518 during training and use images of size 1024×1024 during testing.
3) Vaihingen dataset. The Vaihingen dataset was carried out by the German Association of Photogrammetry andRemote Sensing (Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation, DGPF). The dataset is available at this. It contains 33 tiles with an image resolution of 2400×2400. we crop image patches of size 512×512 from the original tiles in the order from top left to bottom right and add an overlap of 256 pixels when cropping the images. We divide it into a training set (1388 images) and a test set (348 images).

B. Implementation Details

We used NVIDIA GeForce RTX 4090 GPUs during training, with a total of 4 GPUs. The batch size was set to 16. We adopted AdamW as the optimizer to train the model for 50 epochs, and the learning rate was set to $5e^{-6}$ maintaining a constant learning rate throughout the training process without any adjustments. The model was initialized with pre-trained DAM weights. Specifically, we loaded DAM pre-trained weights in each height block in the modified DAM encoder and loaded DAM pre-trained weights in the resolution-agnostic decoder. Since the refine block of DAM does not have multi-scale output, we only loaded DAM pre-trained weights in the last refine block. To prevent overfitting, we performed data augmentation on images and nDSM during training, including horizontal flipping and color distortion with a probability of 0.5. Color distortions include ColorJitter and GaussianBlur. The experiments were conducted using the Pytroch framework.

C. Evaluation Metrics

We use mean absolute error (MAE), root mean squared error (RMSE), and scale invariant root mean squared error (SI_RMSE) to evaluate the effect of model height estimation.

MAE is the average of the absolute values of the differences between the Ground Truth (GT) and the predicted values. It directly measures the average deviation between the predicted values and GT. The MAE formula is as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (7)$$

Compared with MAE, RMSE gives greater weight to larger errors, so RMSE is more sensitive to outliers. The RMSE formula is as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (8)$$

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3564820

Fig. 5. **Qualitative comparison on the GAMUS dataset.** Compared with other methods, Depth2Elevation predicts the height and integrity of buildings more accurately and depicts the subtle height changes of roofs more carefully.

SI_RMSE is a variant of RMSE, which is standardized by dividing by the standard deviation of the true value, making the error metric independent of the scale of the data and eliminating the impact of this scale difference on the evaluation results. The SI_RMSE formula is as follows:

$$\text{SI\_RMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\sigma_y} \tag{9}$$

In the above evaluation metrics formulas, y_i is the GT, $\hat{y}_i$ is the predicted value, n is the number of valid pixels, and σ_y is the standard deviation of the GT. The lower the value of these three indicators, the better the height estimation effect of the model.

TABLE I
QUANTITATIVE RESULTS ON THE GAMUS DATASET. RESOLUTION INDICATES THE SIZE OF THE INPUT IMAGE.

Methods	Resolution	MAE	RMSE	SI_RMSE
IMG2HEIGHT [33]	512×512	4.277	7.348	2.931
HTC-DC [17]	512×512	3.507	6.422	1.221
IMELE [35]	512×512	5.095	7.371	2.617
Depth2Elevation	518×518	**2.135**	**3.701**	**0.802**
IMG2HEIGHT [33]	1024×1024	4.245	7.304	2.413
HTC-DC [17]	1024×1024	2.666	5.200	**0.701**
IMELE [35]	1024×1024	3.720	6.175	1.079
Depth2Elevation	1022×1022	**1.991**	**3.489**	0.986

D. Results

We compare the proposed Depth2Elevation with all open-source single-view image height estimation methods. These methods include IM2HEIGHT [33], IMELE [35], and HTC-DC [17]. We remove the ln of the IMELE loss function L_{depth} and convert it to MAE (mean absolute error) because using ln makes the loss negative. In addition, we do not filter any pixels when calculating the metrics to evaluate the impact of height estimation on the entire image. The results are in meters and all comparison methods are reproduced by us. The best results in each category are in bold.

1) Quantitative comparison with other methods: From Table I, we can see that our proposed Depth2Elevation outperforms the compared methods in all metrics on the GAMUS datasets. Specifically, we use two image sizes of 512×512 and 1024×1024 to train the model, and Depth2Elevation achieves SOTA in the training results of both sizes. It is worth noting that the Depth2Elevation model divides the image into patches of size 14×14, and therefore only accepts images whose dimensions are multiples of 14. To conform to the input image size requirements of Depth2Elevation, when the input image size is 512×512, we resize the image to 518×518.

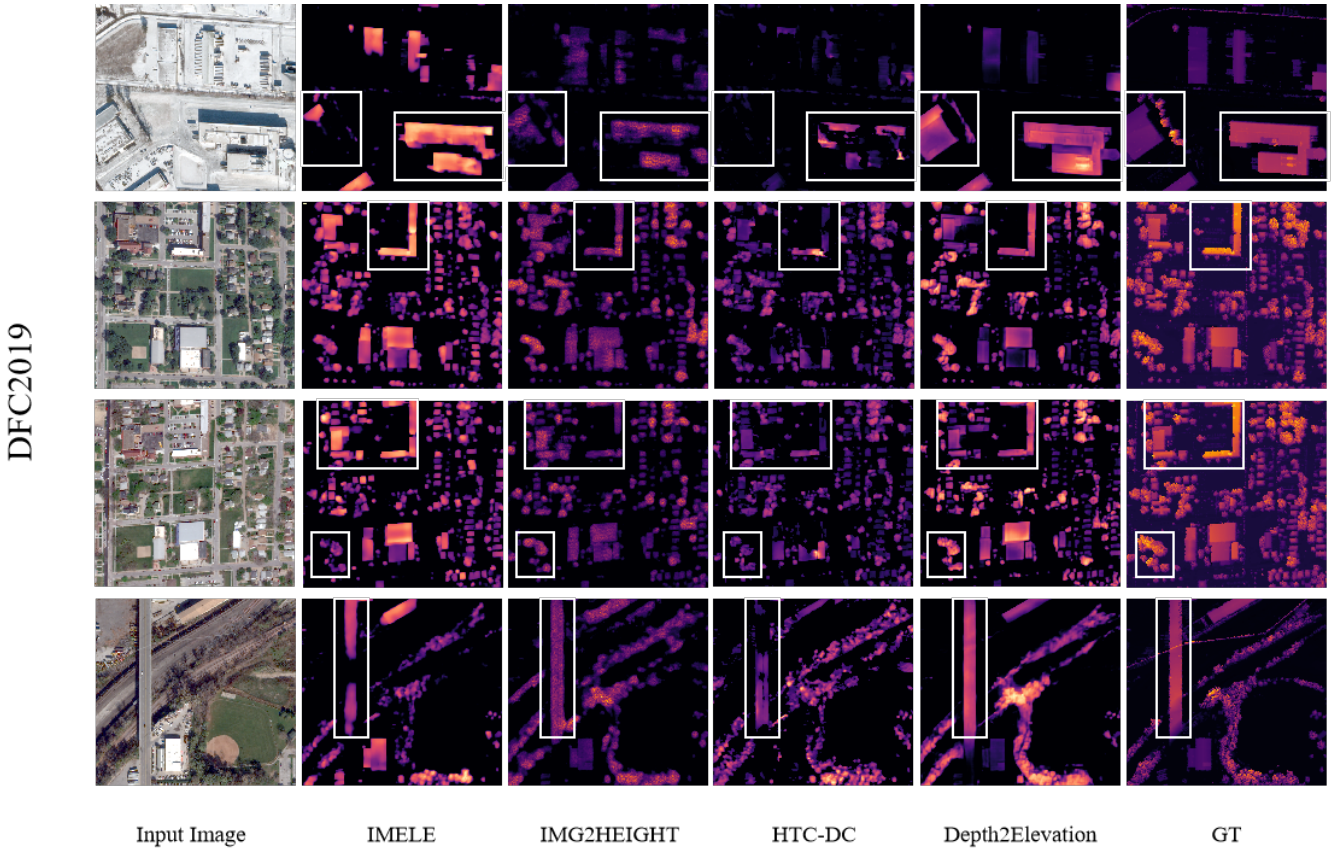

Fig. 6. **Qualitative comparison on the DFC2019 dataset.** Compared with other methods, Depth2Elevation is capable of withstanding color confusion. Even in snowy conditions where the colors of buildings and the ground are similar and there are no significant textural changes, it can still completely predict the height of buildings without any omissions. Additionally, for other objects such as trees and bridges, it can also provide complete height predictions, whereas other methods tend to miss some small trees or provide incomplete predictions for bridges.

TABLE II
QUANTITATIVE RESULTS ON THE DFC2019 DATASET. TRAINING DATASETS REPRESENT THE DATASET USED FOR TRAINING.

Training Datasets	Methods	MAE	RMSE	SI_RMSE
DFC2019	IMG2HEIGHT [33]	1.329	2.728	0.873
	HTC-DC [17]	1.588	3.341	1.036
	IMELE [35]	1.319	2.818	0.865
	Depth2Elevation	**1.138**	**2.450**	**0.749**
GAMUS	IMG2HEIGHT [33]	1.883	3.382	1.069
	HTC-DC [17]	1.611	3.456	1.068
	IMELE [35]	4.356	5.219	1.684
	Depth2Elevation	**1.267**	**2.623**	**0.798**

TABLE III
QUANTITATIVE RESULTS ON THE VAIHINGEN DATASET. THE LABEL SCALE RANGE USED FOR TRAINING IS SCALED TO 0-255.

Datasets	Methods	MAE	RMSE	SI_RMSE
Vaihingen	IMG2HEIGHT [33]	25.899	37.678	0.946
	HTC-DC [17]	17.095	25.831	0.646
	IMELE [35]	15.946	22.410	0.537
	Depth2Elevation	**11.828**	**17.308**	**0.436**

Similarly, when the input image size is 1024×1024, we resize the image to 1022×1022. Among the methods we compared, HTC-DC performs the best. Compared with HTC-DC, Our proposed method performs better by 39.1% (2.135 v.s. 3.507), 42.4% (3.701 v.s. 6.422) and 34.3% (0.802 v.s. 1.221) in terms of MAE, RMSE, and SI_RMSE, respectively. In addition, using 1024×1024 image size for training can achieve better results, because a larger image size contains richer details. When the input image size is 1024×1024, Depth2Elevation is 27.6% (1.991 v.s. 2.666) and 34.9% (3.489 v.s. 5.200) better than HTC-DC in terms of MAE and RMSE metrics. Upon evaluating the models with 1024×1024 resolution imagery, HTC-DC marginally outperforms Depth2Elevation in terms of SI_RMSE. The SI_RMSE serves as a normalized metric for assessing the performance of height estimation models by standardizing the RMSE against the standard deviation of the GT. This approach neutralizes the effects of data scaling, thereby facilitating the comparison of model performance across datasets with varying scales. This superior performance can be attributed to HTC-DC's enhanced accuracy in estimating lower-height values, yet it encounters significant inaccuracies in predicting higher-height values. Conversely, Depth2Elevation maintains a more consistent performance across the entire spectrum of height estimations, which is

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3564820

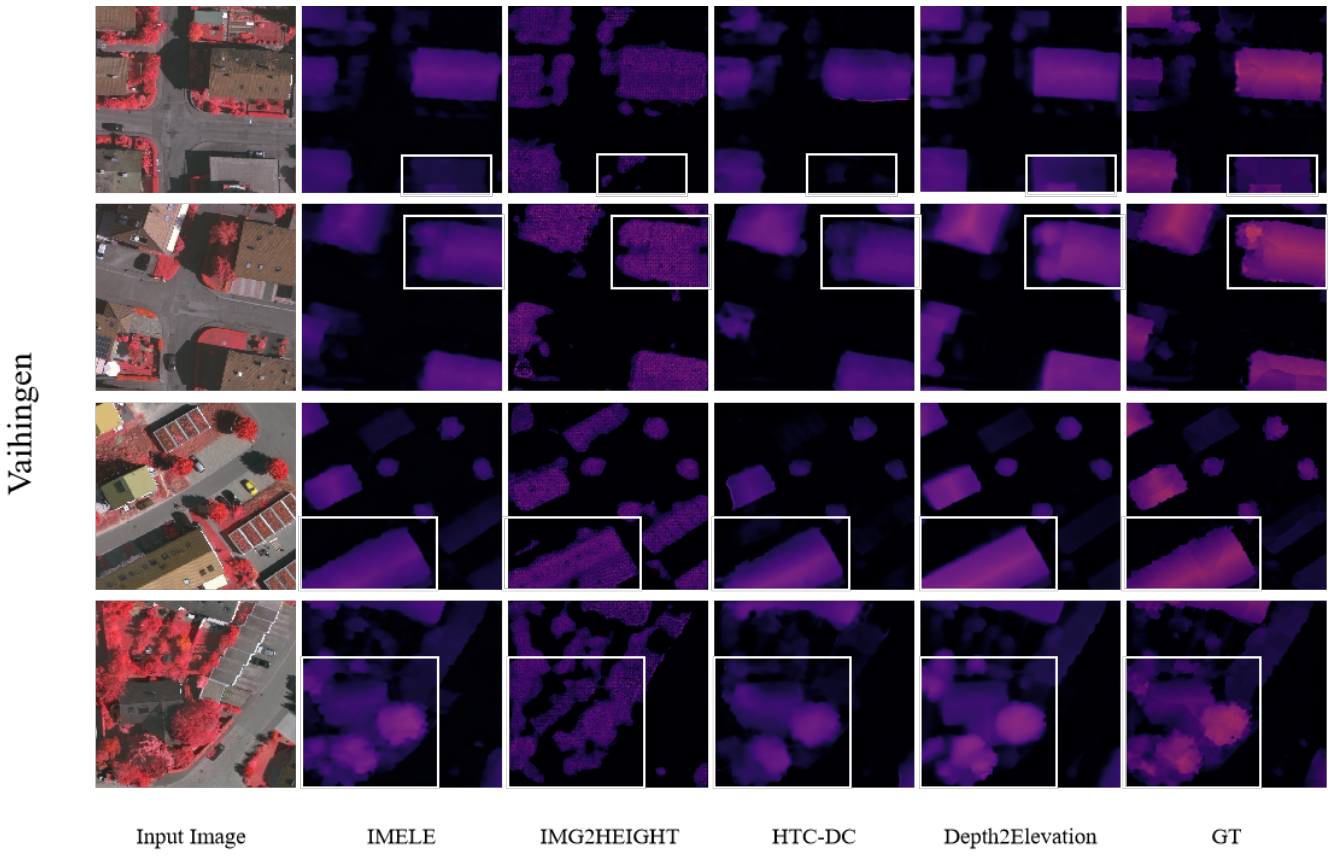

Fig. 7. **Qualitative comparison on the Vaihingen dataset.** Compared with other methods, Depth2Elevation can predict the clear and accurate height boundaries and roof height changes of buildings, and can also more completely predict the height of buildings obscured by trees.

evident in its lower MAE and RMSE compared to HTC-DC. However, when the impact of large-valued errors is normalized in the SI_RMSE metric, Depth2Elevation's performance is slightly surpassed by HTC-DC. This suggests that while HTC-DC may struggle with mitigating extreme prediction errors, Depth2Elevation exhibits a more robust performance across all magnitudes of height estimations.

As shown in the upper part of Table II, Depth2Elevation also achieved SOTA on the DFC2019 dataset. Among the methods we compared, IMELE performs the best. Compared with IMELE, Depth2Elevation exceeds 13.7% (1.138 v.s. 1.319), 13.1% (2.450 v.s. 2.818), and 13.4% (0.749 v.s. 0.865) in MAE, RMSE, and SI_RMSE metrics. Since the images and labels of the DFC2019 dataset and the GAMUS dataset are roughly the same spatial resolution, we tested the generalization of the DFC2019 dataset. We used the model trained on the GAMUS dataset based on 512×512 size images to test directly on the DFC2019 dataset without any fine-tuning. As shown in the lower part of Table II, Depth2Elevation is still the best, which shows its good generalization. Specifically, among the methods we compared, HTC-DC has the best generalization performance, and Depth2Elevation is 21.4% (1.267 v.s. 1.611), 29.1% (2.623 v.s. 3.456), and 25.3% (0.798 v.s. 1.068) better than HTC-DC in MAE, RMSE, and SI_RMSE, respectively.

Table III shows the experimental results based on the Vaihingen dataset, Depth2Elevation is still ahead of other methods. Due to the different label scales of various datasets, the calculated MAE and RMSE evaluation indicators also have scale differences, but SI_RMSE can eliminate scale differences and reflect the accuracy of height estimation. Specifically, among the methods we compared, IMELE performed best on this dataset, and Depth2Elevation is 25.9% (11.828 v.s. 15.946), 22.8% (17.308 v.s. 22.410), and 25.3% (0.436 v.s. 0.537) better than IMELE in MAE, RMSE, and SI_RMSE, respectively.

2) Qualitative comparison with other methods: Fig. 5, Fig. 6 and Fig. 7 shows the visualization results of Depth2Elevation and other comparison methods on GAMUS, DFC2019 and Vaihingen dataset respectively. Color from black to red indicates height from low to high. As can be seen from Fig. 5, on the GAMUS dataset, Depth2Elevation predicts more consistent heights for each object than other methods, and the height prediction for each object more completely describes the object contour. As shown in Fig. 6, on the DFC2019 dataset, Depth2Elevation predicts the height boundary more accurately than other methods, and the height value is closer to the GT. And in Fig. 7, compared with other methods, Depth2Elevation can capture objects missed by other methods and predict the trend of roof height with greater accuracy.

Figure 8 shows the effect of merging the height map predicted by Depth2Elevation and other comparison methods with the input image into a point cloud, which further illustrates the reliability of Depth2Elevation's height prediction. We present a visualization of the GAMUS dataset. In comparison

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3564820

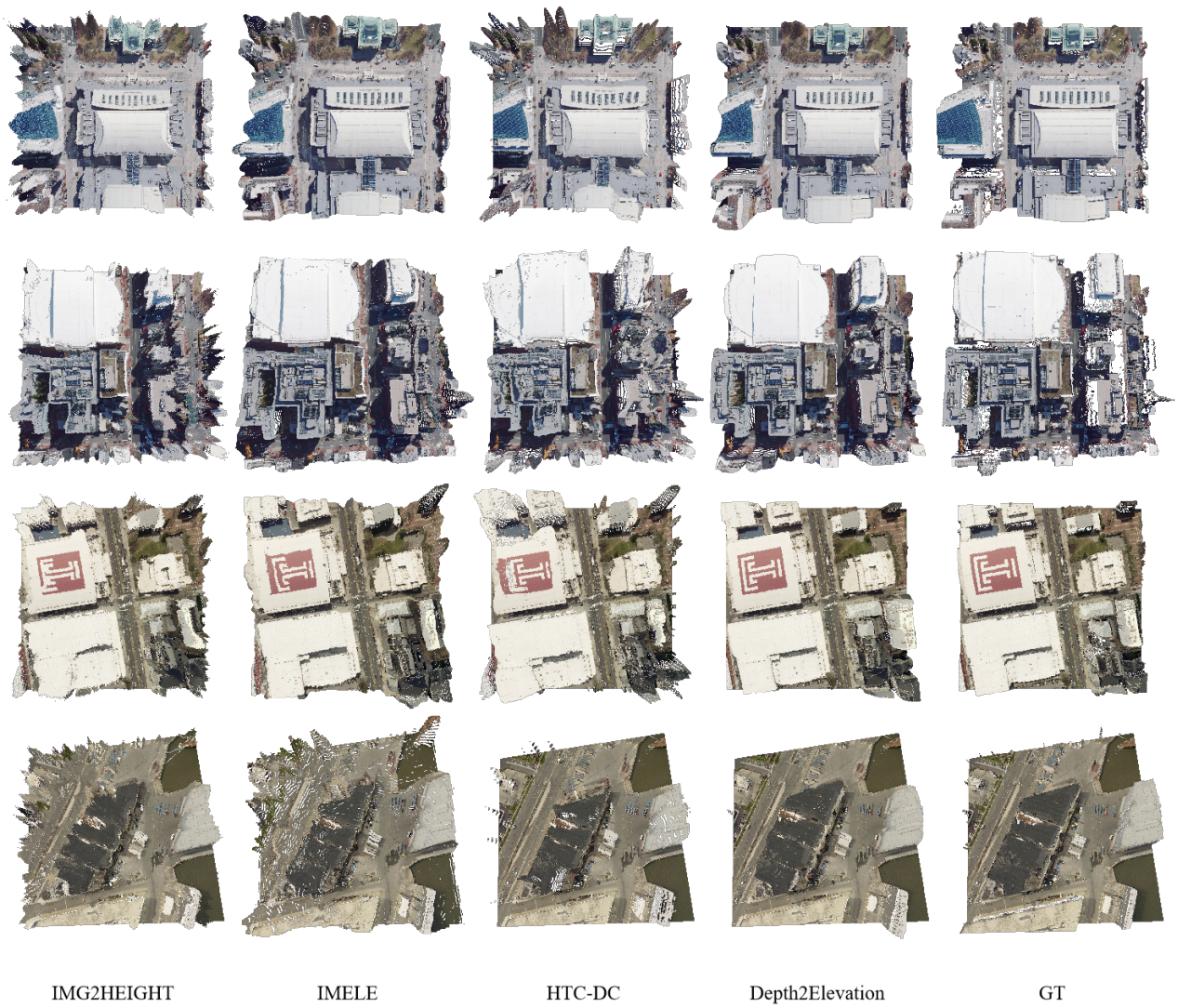

Fig. 8. **Point cloud effect comparison.** The effect of fusing the point cloud formed by the height maps predicted by Depth2Elevation and other comparative methods with the input images. Compared to other methods, Depth2Elevation has less abrupt height at the edge and smoother height predictions on building roofs with fewer outliers.

to other methods, the height predictions from our proposed Depth2Elevation exhibit greater precision, particularly in delineating building contours and capturing local elevations. Conversely, the heights predicted by compared methods result in varying degrees of distortion to the building structures. Collectively, the point cloud generated from the height map predicted by Depth2Elevation more effectively approximates GT, demonstrating its superior performance in height prediction accuracy.

E. Ablation Study

We also performed a thorough ablation study on the various components used in our framework. As a baseline model, we chose the original model of DAM, and similarly used Vit-b as the encoder and frozen it, and the decoder was not modified and fine-tuned. As mentioned in the methodology section, our proposed network has two major enhancements compared to the baseline, namely the use of scale modulator and resolution-agnostic decoder. In order to evaluate how the various components affect the performance, we verified the results of three cases: 1) adding the scale modulator alone; 2) adding the resolution-agnostic decoder alone; 3) adding the multiple loss alone; 4) adding the scale modulator, the resolution-agnostic decoder and the multiple loss simultaneously, that is, Depth2Elevation. All experiments were trained on the GAMUS dataset for 50 epochs, and the evaluation metrics are shown in Table IV. Our experiment uses images of size 518×518 as input for training, and the image size is 1024×1024 during testing.

Based on the experimental results, it is evident that the introduction of the scale modulator, the resolution-agnostic decoder, and the multiple loss have all led to significant enhancements compared to the baseline model. This indicates that each of these components, the scale modulator, the resolution-agnostic decoder, and the multiple loss, is highly effective, and their combination exerts an even greater impact. Specifically, the addition of the scale modulator results in a notable decrease

TABLE IV
ABLATION STUDY ON THE GAMUS DATASET. RESULTS ARE COMPARED IN METERS. THE TRAINING IMAGE SIZE IS 518×518.

Modules	MAE	RMSE	SI_RMSE
baseline [25]	3.171	5.343	1.784
baseline + scale modulator	2.354	3.918	1.076
baseline + resolution-agnostic decoder	2.632	3.950	2.313
baseline + multiple loss	2.663	4.364	2.206
Depth2Elevation	**2.135**	**3.701**	**0.802**

TABLE V
QUANTITATIVE COMPARISON OF DEPTH2ELEVATION AND DIRECT FINE-TUNING OF DAM. TRAINING DATASETS REPRESENT THE DATASET USED FOR TRAINING. DAM REFERS TO DIRECT FINE-TUNING OF DAM.

Training Datasets	Methods	MAE	RMSE	SI_RMSE
GAMUS	DAM [25]	3.171	5.343	1.784
	Depth2Elevation	**1.991**	**3.489**	**0.986**
DFC2019	DAM [25]	1.537	2.603	0.809
	Depth2Elevation	**1.138**	**2.450**	**0.749**
Vaihingen	DAM [25]	13.954	19.891	0.500
	Depth2Elevation	**11.828**	**17.308**	**0.436**

in MAE, RMSE, and SI_RMSE, demonstrating its effectiveness in enhancing the model's ability to capture features across different scales. The incorporation of the Resolution-Agnostic Decoder leads to a slight decrease in MAE, RMSE, and SI_RMSE, indicating that it better integrates multi-scale features when processing images of varying resolutions. The implementation of the multiple loss significantly reduces MAE and RMSE, with an improvement also observed in SI_RMSE, suggesting that it effectively enhances the model's capacity to capture geometric structures and mitigates the influence of extreme values and outliers. When the scale modulator, the resolution-agnostic decoder, and the multiple loss are combined, the model's performance markedly improves, with MAE, RMSE, and SI_RMSE all reaching optimal levels. This underscores the critical role of the synergistic interaction of these components in elevating model performance. The outcomes of the ablation studies demonstrate that the scale modulator, the resolution-agnostic decoder, and the multiple loss play a significant role in enhancing the performance of the Depth2Elevation model. The synergistic effect of these components enables the model to more accurately capture features at different scales, thereby improving the precision and robustness of height estimation.

V. DISCUSSION

A. Benefits of Scale Modulation and resolution-agnostic decoder Fine-tuning of DAM

Based on this study, this section provides a comprehensive and detailed analysis of the advantages of fine-tuning the foundation model DAM using scale modulation and resolution-agnostic decoder compared to direct fine-tuning of the DAM. Direct fine-tuning means freezing the encoder of DAM and only training the decoder.

Table V presents the test results of the proposed method Depth2Elevation and direct fine-tuning of DAM on the GAMUS, DFC2019, and Vaihingen datasets. It can be observed from the table that the results of Depth2Elevation are superior to those of direct fine-tuning of DAM, confirming the effectiveness of fine-tuning based on scale modulation and a resolution-agnostic decoder.

Fig. 9 illustrates the height prediction effects of this study and direct fine-tuning of DAM on the GAMUS, DFC2019, and Vaihingen datasets, where the color gradient from black to red represents increasing height. It can be seen that in the smaller-scale Vaihingen dataset, direct fine-tuning of DAM can achieve results close to those of Depth2Elevation. However, when predicting larger-scale datasets such as GAMUS and DFC2019, direct fine-tuning of DAM produces a significant amount of artifacts in the height maps, with inaccurate predicted boundaries and overall blurriness. In contrast, Depth2Elevation can generate precise and clear height maps. This indicates that the proposed fine-tuning based on scale modulation and multi-scale is better suited to handle remote sensing images of different scales, capturing the geometric structure of objects of various scales and sizes more effectively, and can efficiently transfer the foundation model DAM to the task of single-view remote sensing image height estimation.

B. The Computational Complexity and Efficiency of Fine-tuning DAM with Scale Modulation and resolution-agnostic decoder

In this section, we assessed the computational complexity and efficiency of Depth2Elevation compared to other methods. We conducted experiments on an NVIDIA GeForce RTX 4090 GPU to measure both computational complexity and inference time.

Computational complexity is an important metric for gauging the efficiency of an algorithm, involving the computational resources required during its execution. Parameters and floating-point operations (FLOPs) are two key standards for measurement. Depth2Elevation has $99.540M$ parameters, which is moderate compared to IMELE's $158.389M$ and HTC-DC's $51.634M$. This indicates that Depth2Elevation is reasonable in terms of model complexity, not excessively increasing the number of parameters in pursuit of performance, yet still achieving state-of-the-art (SOTA) results. The FLOPs of Depth2Elevation is $1406.583G$, lower than IMELE's $4250.403G$ and HTC-DC's $1615.929G$. This suggests that Depth2Elevation maintains high performance without significant computational complexity.

Inference time is another crucial metric for evaluating the practicality of a model. The inference time of Depth2Elevation is $12.966ms$, faster than IMELE's $45.805ms$ and HTC-DC's $36.083ms$. This result indicates that Depth2Elevation can provide quicker responses in practical applications.

Overall, Depth2Elevation effectively controls computational complexity and inference time while maintaining high performance. It ensures model performance without increasing the demand for computational resources. This balance allows Depth2Elevation to achieve SOTA results without the need for extensive computational resources.

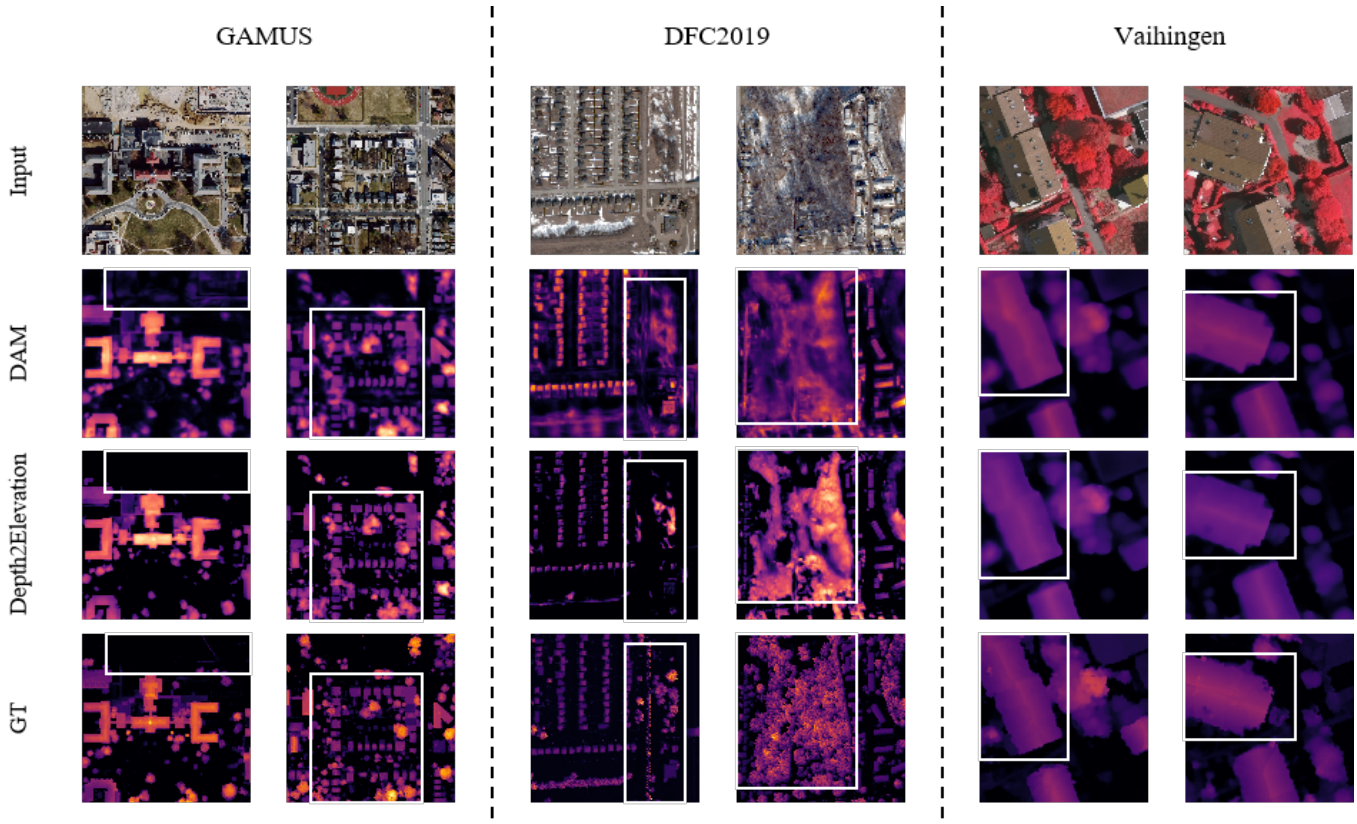

Fig. 9. **Comparison of height maps predicted by Depth2Elevation and directly fine-tuned DAM.** Compared with directly fine-tuning DAM, Depth2Elevation has no artifacts on the ground and predicts the height change of the roof more accurately.

TABLE VI
COMPUTATIONAL COMPLEXITY AND INFERENCE TIME STATISTICS OF DEPTH2ELEVATION AND OTHER COMPARED METHODS. RESOLUTION REFERS TO THE SIZE OF THE IMAGE BEING FED INTO THE MODEL.

Methods	Resolution	Parameters(M)	FLOPs(G)	Inference Time(ms)
IMG2HEIGHT [33]	1024 × 1024	7.360	1003.886	6.448
IMELE [35]	1024 × 1024	158.389	4250.403	45.805
HTC-DC [17]	1024 × 1024	51.634	1615.929	36.083
Depth2Elevation	1036 × 1036	99.540	1406.583	12.966

[1] We calculate parameters and FLOPs by THOP.

C. Visualization of Height Map Errors Predicted by Depth2Elevation and Compared Methods

We calculated errors between height maps predicted by Depth2Elevation, Directly Fine-tuned DAM, and Compared Methods and GT, and visualize these errors, which are displayed in Fig. 10. In the process of generating error maps, we first compute the absolute differences between predicted depth maps and GT, and then normalize these differences to calculate error percentages. Subsequently, based on the normalized error values for each pixel, we map them to corresponding colors to create error images. To facilitate identification, a set of color labels was added to upper left corner of error images, with each label corresponding to a specific error range and color value. The error maps can intuitively present distribution of errors between predicted depth maps and GT. By observing the error maps, we can visually see where the models have higher prediction accuracy and where there are larger errors.

Specifically, we analyze Fig. 10 to assess the performance of different models. On the GAMUS dataset, IMELE and Directly Fine-tuned DAM exhibit significant errors in ground areas, indicating the presence of artifacts in their predicted height maps, and they are unable to accurately distinguish between the ground and objects on the ground. Compared to IMG2HEIGHT and HTC-DC, Depth2Elevation show light blue areas in the error maps for parts of the buildings, suggesting that local height predictions for targets have fewer errors. This indicates that Depth2Elevation is better at capturing subtle changes in object heights. In the DFC2019 dataset, the DAM Directly Fine-tuned still shows extensive errors in ground height predictions, while other methods have more blue areas compared to Depth2Elevation. Additionally, IMELE and HTC-DC exhibit noticeable errors in the ground areas between two rows of houses while Depth2Elevation

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2025.3564820

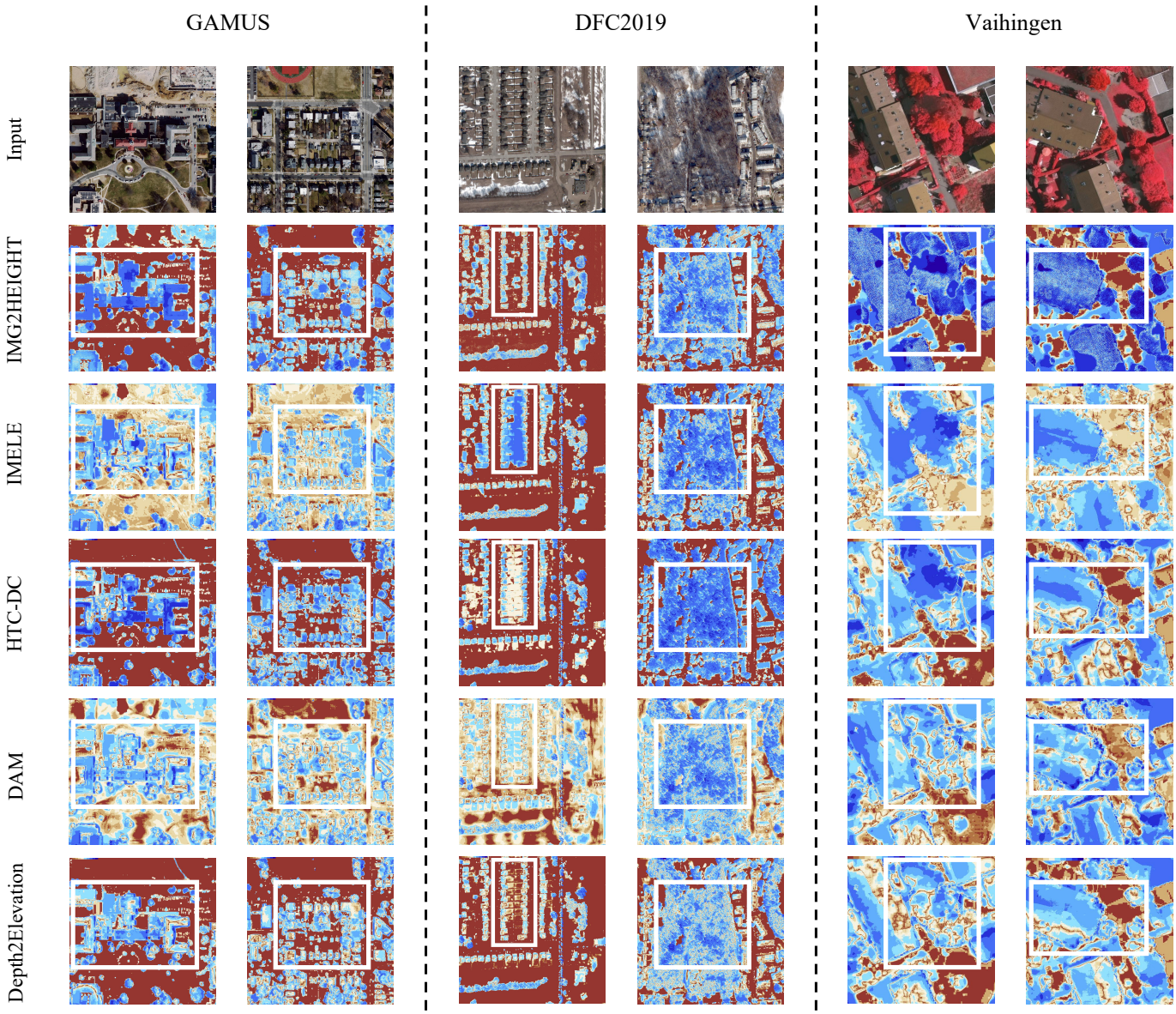

Fig. 10. **Error map effect comparison.** Calculated error maps by Depth2Elevation, direct fine-tuning of DAM, and compared methods. The error value increases from red to blue.

has fewer, indicating that Depth2Elevation is more accurate in capturing the boundaries of objects. On the Vaihingen dataset, IMELE has significant errors in the ground areas, while Depth2Elevation has fewer errors in the ground areas and even smaller errors in the building roof areas compared to other methods, as indicated by the lighter blue color. In general, our proposed Depth2Elevation can accurately predict the contours and boundaries of the objects and capture the undulations in the heights of the objects, thus achieving more accurate height predictions.

D. Limitations and Future Work

Although the Depth2Elevation model has shown impressive performance on several benchmark datasets, there are still some limitations that need to be further improved in future research.

1) Limited generalization ability. The performance of the Depth2Elevation model on cross-domain datasets still needs to be improved. For example, on the OsiDataset [35] with lower spatial resolution, the model's performance drops significantly. This indicates that the current model is relatively sensitive to data distribution, especially when facing data from different sensor types or geographical regions, its generalization ability is limited. Future research can explore more powerful data augmentation strategies, transfer learning methods, or meta-learning techniques to enhance the model's adaptability and robustness across different datasets.
2) Computational Efficiency. The number of parameters of the Depth2Elevation model is close to that of HTC-DC and much larger than that of IMG2HEIGHT. This may limit its application on resource-constrained devices,

such as edge computing devices or mobile devices. To improve the model's practicality, future research can explore more lightweight network structures, such as reducing the number of parameters and computational complexity through model compression, knowledge distillation, or designing more efficient network architectures, while maintaining or improving the model's performance.

VI. Conclusion

In this study, we propose the Depth2Elevation model to estimate the height of a single-view remote sensing image, which is the first to use DAM for single-view height estimation. We fine-tune the DAM using scale modulation and develop a resolution-agnostic decoder to enhance the model's ability to extract features of different scales, enabling the model to learn semantic and fine-grained geometric information simultaneously. We conclude that our strategy of transferring DAM to remote sensing height estimation tasks has better results and generalization than traditional end-to-end model training. In the future, we will explore more efficient fine-tuning methods and transfer the Depth2Elevation model to different domain datasets.

Acknowledgements

This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (No. 42401567) and the Tertiary Education Scientific research project of Guangzhou Municipal Education Bureau (No. 2024312159). The authors would also like to thank Sining Chen for helping debug the code of HTC-DC [17].

References

[1] Sagar S Deshpande. Bank line extraction by integration of orthoimages and lidar digital elevation model using principal component analysis and alpha matting. *Photogrammetric Engineering & Remote Sensing*, 90(10):631–638, 2024.

[2] Ismael Colomina and Pere Molina. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS Journal of photogrammetry and remote sensing*, 92:79–97, 2014.

[3] Guozhang Liu, Baochai Peng, Ting Liu, Pan Zhang, Mengke Yuan, Chaoran Lu, Ningning Cao, Sen Zhang, Simin Huang, Tao Wang, Xiaoqiang Lu, Licheng Jiao, Qiong Liu, Lingling Li, Fang Liu, Xu Liu, Yuting Yang, Kaiqiang Chen, Zhiyuan Yan, Deke Tang, Hai Huang, Michael Schmitt, Xian Sun, Gemine Vivone, Claudio Persello, and Ronny Hänsch. Large-scale fine-grained building classification and height estimation for semantic urban reconstruction: Outcome of the 2023 ieee grss data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:11194–11207, 2024.

[4] Marcela Carvalho, Bertrand Le Saux, Pauline Trouvé-Peloux, Frédéric Champagnat, and Andrés Almansa. Multitask learning of height and semantics from aerial images. *IEEE Geoscience and Remote Sensing Letters*, 17(8):1391–1395, 2020.

[5] Jie Dong, Wufan Zhao, and Shuai Wang. Multiscale context aggregation network for building change detection using high resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[6] Milad Salehi-Dorcheabedi, Jamal Asgari, Alireza Amiri-Simkooei, and Sayyed Bagher Fatemi Nasrabadi. Improving lidar height precision in urban environment: Low-cost gnss ranging prototype for post-mission airborne laser scanning enhancement. *Remote Sensing Applications: Society and Environment*, 35:101251, 2024.

[7] Renato Juliano Martins, Emil Marinov, M Aziz Ben Youssef, Christina Kyrou, Mathilde Joubert, Constance Colmagro, Valentin Gâté, Colette Turbil, Pierre-Marie Coulon, Daniel Turover, et al. Metasurface-enhanced light detection and ranging technology. *Nature Communications*, 13(1):5724, 2022.

[8] Aire Olesk, Kaupo Voormansik, Ants Vain, Mart Noorma, and Jaan Praks. Seasonal differences in forest height estimation from interferometric tandem-x coherence data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(12):5565–5572, 2015.

[9] Shouhang Du, Jianghe Xing, Shaoyu Wang, Xiongwu Xiao, Jun Li, and Hao Liu. Lumnet: Land use knowledge guided multiscale network for height estimation from single remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 2024.

[10] Saket Kunwar. U-net ensemble for semantic and height estimation using coarse-map initialization. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 4959–4962, 2019.

[11] Zhuo Zheng, Yanfei Zhong, and Junjue Wang. Popnet: Encoder-dual decoder for semantic segmentation and single-view height estimation. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 4963–4966, 2019.

[12] Wufan Zhao, Hu Ding, Jiaming Na, Mengmeng Li, and Dirk Tiede. Height estimation from single aerial imagery using contrastive learning based multi-scale refinement network. *International Journal of Digital Earth*, 16(1):2322–2340, 2023.

[13] Xiang Li, Mingyang Wang, and Yi Fang. Height estimation from single aerial images using a deep ordinal regression network. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2020.

[14] Siyuan Xing, Qiulei Dong, and Zhanyi Hu. Gated feature aggregation for height estimation from single aerial images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[15] Wufan Zhao, Claudio Persello, and Alfred Stein. Semantic-aware unsupervised domain adaptation for height estimation from single-view aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:372–385, 2023.

[16] Zhitong Xiong, Sining Chen, Yilei Shi, and Xiao Xiang Zhu. Self-supervised pre-training with monocular height

estimation for semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[17] Sining Chen, Yilei Shi, Zhitong Xiong, and Xiao Xiang Zhu. Htc-dc net: Monocular height estimation from single remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–18, 2023.

[18] Zhitong Xiong, Wei Huang, Jingtao Hu, and Xiao Xiang Zhu. The benchmark: Transferable representation learning for monocular height estimation. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

[19] Jiangyan Wu, Mengke Yuan, Tong Wang, Xiaohong Jia, and Dong-Ming Yan. Heightformer: Single imagery height estimation transformer with bilateral feature pyramid fusion. *IEEE Geoscience and Remote Sensing Letters*, 2024.

[20] Zhan Chen, Yidan Zhang, Xiyu Qi, Yongqiang Mao, Xin Zhou, Lei Wang, and Yunping Ge. Heightformer: A multilevel interaction and image-adaptive classification–regression network for monocular height estimation with aerial images. *Remote Sensing*, 16(2):295, 2024.

[21] Siyuan Wang, Bowen Cai, Dongyang Hou, Qing Ding, Jiaming Wang, and Zhenfeng Shao. Mf-bhnet: A hybrid multimodal fusion network for building height estimation using sentinel-1 and sentinel-2 imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–19, 2024.

[22] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27672–27683, 2024.

[23] Xiaoyan Lu and Qihao Weng. Multi-lora fine-tuned segment anything model for urban man-made object extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–19, 2024.

[24] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2022.

[25] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024.

[26] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

[27] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

[28] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[29] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4009–4018, 2021.

[30] Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, and Zhengguo Li. Nddepth: Normal-distance assisted monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7931–7940, 2023.

[31] Xiaodong Yang, Zhuang Ma, Zhiyu Ji, and Zhe Ren. Gedepth: Ground embedding for monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12719–12727, 2023.

[32] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024.

[33] Lichao Mou and Xiao Xiang Zhu. Im2height: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network. *arXiv preprint arXiv:1802.10249*, 2018.

[34] Savvas Karatsiolis, Andreas Kamilaris, and Ian Cole. Img2ndsm: Height estimation from single airborne rgb images with deep learning. *Remote Sensing*, 13(12):2417, 2021.

[35] Chao-Jung Liu, Vladimir A Krylov, Paul Kane, Geraldine Kavanagh, and Rozenn Dahyot. Im2elevation: Building height estimation from single-view aerial imagery. *remote sensing*, 12(17):2719, 2020.

[36] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *CoRR*, abs/2304.07193, 2023.

[37] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

[38] Zhitong Xiong, Sining Chen, Yi Wang, Lichao Mou, and Xiao Xiang Zhu. Gamus: A geometry-aware multi-modal semantic segmentation benchmark for remote sensing data. *arXiv preprint arXiv:2305.14914*, 2023.

[39] Bertrand Le Saux, Naoto Yokoya, Ronny Hänsch, and Myron Brown. Data fusion contest 2019 (dfc2019), 2019.