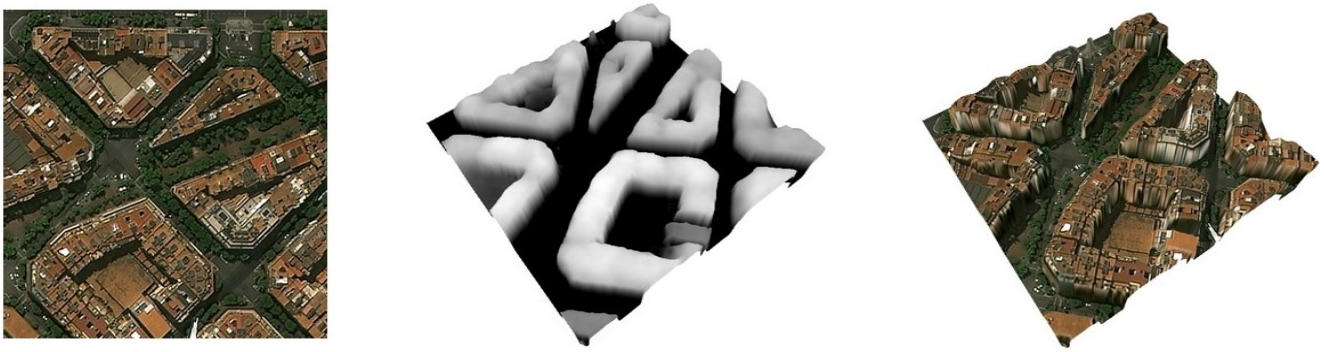
Fusing Convolution and Vision Transformer Encoders for Object Height Estimation from Monocular Satellite and Aerial Images

Furkan Gültekin[1] Alper Koz[3] Reza Bahmanyar[2] Seyed Majid Azimi[2] Mehmet Lütfi Süzen[4]
[1]Locdus AI [2]German Aerospace Center (DLR), Wessling, Germany
[3] Center for Image Analysis, METU [4]Geological Engineering Department, METU



3D reconstruction of a satellite image using the predicted height by the proposed FusedSeg-HE model

Abstract

Accurate height estimation from aerial and satellite imagery is crucial for large-scale 3D scene modeling, which has applications in urban planning, environmental monitoring, and disaster management. In this work, we propose integrating convolutional neural networks (CNNs) and vision transformers (ViTs) to leverage both local and global feature extraction. Our experiments show that using a combination of CNN and ViT encoders significantly improves accuracy compared to relying on either one alone, as CNNs capture fine details while ViTs enhance contextual understanding. Additionally, we incorporate a segmentation head to enhance pixel-level precision, particularly at object boundaries. Evaluated on the DFC2019 and DFC2023 datasets, our proposed fusion approach outperforms baseline methods across multiple metrics. For instance, root-mean-squared error is reduced by 5%–13%, and accuracy is improved by 4%–9% in the delta threshold metric. The results also demonstrate strong generalizability across diverse sensors, acquisition altitudes, viewing angles, and real-world scenarios. Our models are released at https://github.com/Furkangultekin/FusedHE.

1. Introduction

The extraction of height information from satellite and aerial images is conventionally achieved using photogrammetric methods [6, 30]. However, these methods require human supervision and multiple images from different angles, and are also dependent on various parameters, such as camera settings and flight altitudes. Recently, advancements in artificial intelligence and the availability of labelled datasets have driven interest in deep learning-based approaches. These approaches can extract depth or height information even from a single image without manual intervention or additional parameter tuning.

Obtaining meaningful elevation information only from pixel brightness in optical imagery is challenging; therefore, both local and global features must be considered. As illustrated in Fig. 1, in the areas indicated in the blue and green boxes, pixels that are close to each other can carry information from different objects and represent different local features. Conversely, pixels distant from each other in the areas indicated in the red and yellow boxes can carry information from a similar object and reveal global pixel relationships.

Convolutional Neural Networks (CNNs) [16, 18, 23, 24, 36, 37] and Vision Transformers (ViTs) [10, 11, 28, 39, 40, 42, 43] are two prominent deep learning architectures frequently utilized for in-situ image depth estimation. These architectures have been adapted for height estimation tasks in satellite imagery. CNNs focus on extracting local

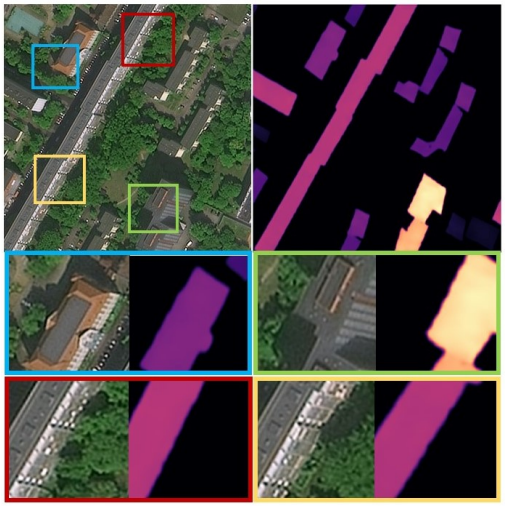


Figure 1. A satellite imagery with its height map and zoomed-in examples demonstrate, local features can provide information about different objects in neighboring pixels (blue and green examples), while distant features can carry information about similar objects (red and yellow examples).

features through convolutional layers and capturing spatial feature hierarchies. In contrast, ViTs utilize self-attention mechanisms to model long-range dependencies and global context, achieving broader image understanding. The existing methods for height estimation from single images, whether employing CNNs or ViTs, demonstrate significant limitations, emphasizing the complexity of extracting and effectively utilizing all relevant local and global features from a single image. This underscores the necessity to leverage the complementary strengths of these models.

To address this need, we propose an encoder-level general fusion strategy that combines the strengths of CNNs and ViTs for height estimation from monocular satellite images. This Fused Height Estimation (Fused-HE) method operates by extracting features from two parallel encoder branches—one CNN-based and one ViT-based—and fusing them at multiple scales through a series of Feature Fusion Blocks (FFBs). These fused features are then passed to a shared decoder with skip connections to retain spatial detail and enhance prediction accuracy. We evaluated the Fused-HE method using multiple encoder combinations, including CNNs, standard ViTs, and hierarchical ViTs. The results show that fusing ResNet-101 [18] with standard ViTs, such as DPT [34], marginally improves performance. However, combining ResNet-101 with hierarchical ViTs like Mix Transformer Encoder (MiT)-B4 [43] significantly improves accuracy, highlighting the effectiveness of integrating local and multi-scale global features. These results imply that encoder-level fusion success depends not only on the fusion mechanism itself, but also on the compatibility between the CNN and ViT backbones.

To improve the accuracy of height estimation at the pixel level, we incorporate a segmentation head into the architecture. This head uses features from the shared encoder-decoder network to predict object masks. This enables the allocation of height values to the appropriate pixels. We call this method Fused Segmentation Height Estimation (FusedSeg-HE). We evaluated the Fused-HE method with and without auxiliary segmentation, as well as multiple CNN-ViT encoder combinations, against baseline single-encoder models on the DFC2019 [35] and DFC2023 [32] datasets. The baseline models include: **CNN-HE**, which uses ResNet-101 and excels in local feature extraction, particularly capturing sharp object boundaries; **ViT-HE**, based on the DPT-Base encoder, which focuses on global contextual understanding but often lacks precision at object boundaries; and **MiT-HE**, which leverages the hierarchical MiT-B4 encoder for multi-scale feature extraction, offering a balance between detail and generalization.

The evaluations concluded that: **i)** Fused-HE-based configurations employing convolutional and hierarchical vision transformer encoders consistently achieved superior performance in both quantitative and qualitative assessments, confirming the effectiveness of jointly extracting local and global features through encoder-level fusion; **ii)** the inclusion of a segmentation-aware auxiliary head, as implemented in the FusedSeg-HE variant, provided additional gains in pixel-level accuracy—particularly at object boundaries—demonstrating the benefits of incorporating semantic guidance; **iii)** training on the DFC2023 dataset, which offers global coverage and strong diversity, enables the models to achieve superior generalization, reinforcing its robustness across varying geographic and structural contexts; **iv)** validations on images from various satellite and airborne platforms demonstrate the generalizability of our proposed fusion methods for real-world applications.

2. Related Work

The success of **CNN-based** models, such as the Multi-Scale Deep Network [12] and MiDaS [33] in in-situ imagery, has led to the increased popularity of CNNs for height estimation in satellite and aerial imagery. The models IM2height [31], IM2Elevation [27], and IMG2nDSM [20] have demonstrated the effectiveness of multi-scale feature extraction, showing promising results. The incorporation of hierarchical architectures, such as ResNet, into these models serves to further enhance their capabilities by facilitating the capture of both local and high-level image features [1, 25].

ViTs are also demonstrating success in depth estimation tasks on in-situ images, with models like Dense Prediction Transformers (DPT) [34], GLP-Depth [21], and Depth Anything [46] combining ViTs with feature extraction for precise depth predictions. Other approaches, such

as AdaBins [5] and BinsFormer [26], improve performance by discretizing depth values into bins and estimating their center values. However the application of ViTs in height estimation from single aerial and satellite images is limited. The Knowledge Transfer for Label-Efficient Monocular Height Estimation model [45] uses Swin transformers for feature extraction, leveraging pre-trained synthetic data for transfer learning before fine-tuning on real satellite images. Similarly, in [8, 9], the authors integrate ViTs into a CNN-based encoder-decoder network, transforming height estimation into a classification-regression task using AdaBins.

Studies **combining monocular depth estimation with semantic segmentation** have shown improved accuracy in both tasks on in-situ images. The authors in [17] used a separate encoder-decoder model for segmentation guidance in self-supervised depth estimation. The authors in [22] employed a shared encoder with two separate decoders for segmentation and depth estimation. Several recent works [7, 41, 44] have approached height estimation and semantic segmentation as a multi-task learning problems in the context of aerial and satellite images. These models employ a single CNN encoder, such as ResNet, to extract shared feature representations, which are then passed through multiple task-specific decoders to produce both height maps and semantic segmentation outputs.

In **supervised learning**, models are trained using annotated ground truth data, whereas in **self-supervised learning**, they learn directly from input data without labeled supervision. Self-supervised methods are widely used in monocular depth estimation [13, 14, 29, 48], with models like MonoDepth2 [15] leveraging convolutional encoders for depth prediction and auxiliary networks for pose estimation. Recently, self-supervised ViTs such as MonoFormer [4] and MonoViT [47] have been proposed, combining transformers with CNNs for enhanced feature extraction. Although self-supervised methods effectively estimate relative depth by extracting disparities, they struggle with absolute height prediction, particularly in satellite imagery, due to the lack of ground truth data. This limitation makes supervised methods the more suitable choice for the height estimation tasks in this study.

3. Dataset

In this work, we use the DFC2019 [35] and DFC2023 [32] datasets, published as part of the annual Data Fusion Contest (DFC) organized by the Institute of Electrical and Electronics Engineers (IEEE). These datasets contain satellite images along with corresponding height data in the form of nDSM as ground truth, enabling the training of models to predict height values from single satellite images.

The **DFC2019** dataset consists of 26 WorldView(WV)-3 satellite images of Jacksonville, Florida, captured between 2014 and 2016, and 43 images of Omaha, Nebraska, captured between 2014 and 2015. It includes panchromatic, 8-band visible, and near-infrared (VNIR) images. The Ground Sampling Distance (GSD) is approximately 35cm per pixel for panchromatic images and 1.35m for 8-band visible and VNIR images, all of which are pan-sharpened. The nDSM raster data is derived from Airborne LiDAR, with an Aggregate Nominal Pulse Spacing (ANPS) of approximately 80cm. The **DFC2023** dataset includes a large collection of optical images from the SuperView-1, Gaofen-2, and Gaofen-3 satellites, with GSD of 0.5m, 0.8m, and 1m, respectively. The nDSM raster data is generated from stereo images captured by the Gaofen-7, WV-1 and -2 satellites, with a GSD of roughly 2m. The dataset covers images from seventeen cities across six continents. Details on the datasets are in the supplementary materials.

4. Methods

This section explains our proposed encoder fusion method, Fused-HE, and its variant with an additional segmentation head, FusedSeg-HE.

4.1. Fused-HE

In principle, Fused-HE is an encoder-decoder supervised learning method in which a CNN and a ViT are fused to serve as the encoder. For the explanation, we consider ResNet-101 and MiT-B4 as the fused CNN and hierarchical ViT encoders, respectively. Fig. 2 illustrates an overview of the method. As shown, the input satellite image is processed independently by the two encoders, and their multi-scale features are aligned and fused using Feature Fusion Blocks (FFBs). The fused outputs are then decoded via skip connections to refine spatial detail. Finally, the height map is resized to match the input dimensions.

CNN encoder: The ResNet-101 encoder consists of four main convolutional blocks, each comprising three convolution layers: $\text{Conv}_{1\times1}$, $\text{Conv}_{3\times3}$, and $\text{Conv}_{1\times1}$. These layers are repeated 3, 4, 23, and 3 times in the respective blocks. The output size of each block $i = \{1, 2, 3, 4\}$ is given by: $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_{\text{resnet}_i}$, where H and W denote the height and width of the input image, respectively, and $C_{\text{resnet}_i} \in \{256, 512, 1024, 2048\}$.

ViT encoder: The MiT-B4 encoder integrates global feature extraction with a hierarchical CNN-like structure. Unlike standard ViTs, it applies overlapped patch merging at the end of each transformer block, dynamically adjusting feature sizes within the encoder. As illustrated in Fig. 2, it consists of four main transformer blocks, each generating output at different resolutions through a sequence of self-attention and Mix-FFN, a feedforward layer introduced in [43]. The number of repetitions for each block is 3, 8, 27, and 3, respectively. At the end of each block, an overlapped patch merging unit is applied. The output resolution of each

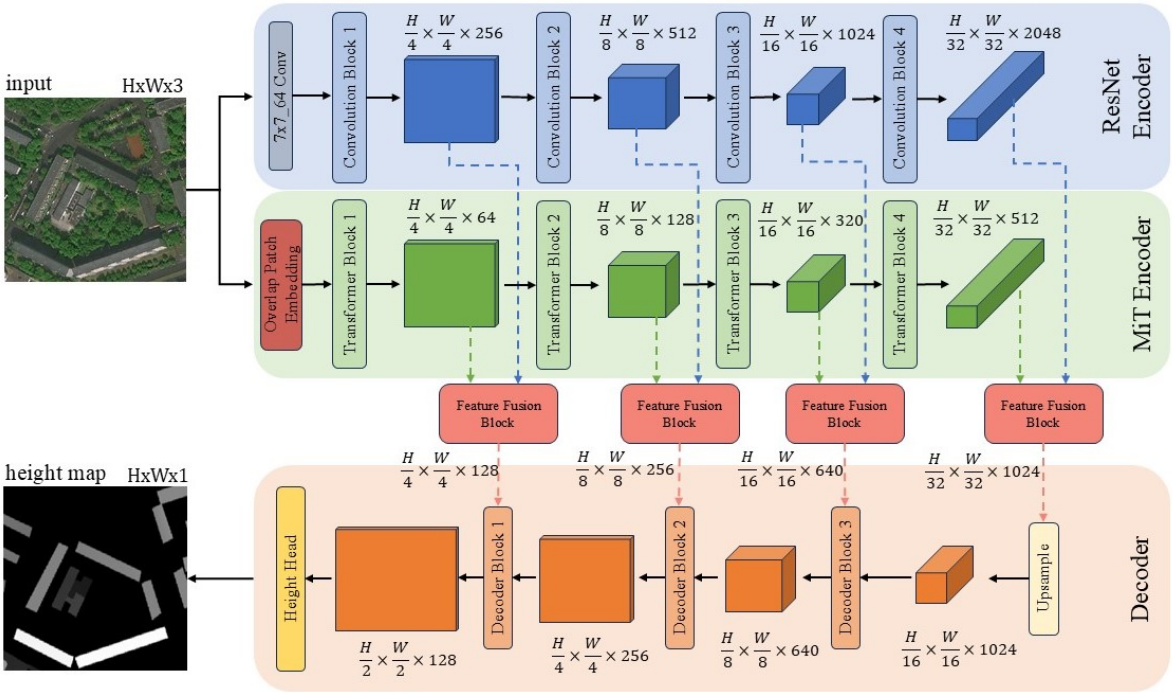


Figure 2. Fused-HE method with ResNet-MiT encoder. The input image first goes through the Resnet Encoder and MiT Encoder separately in the upper row, where the resulting features after the convolution and transformer block at each scale are concatenated with the features fusion block. The outputs of the fusing blocks are successively decoded at each scale at the decoder side, illustrated in the lower row.

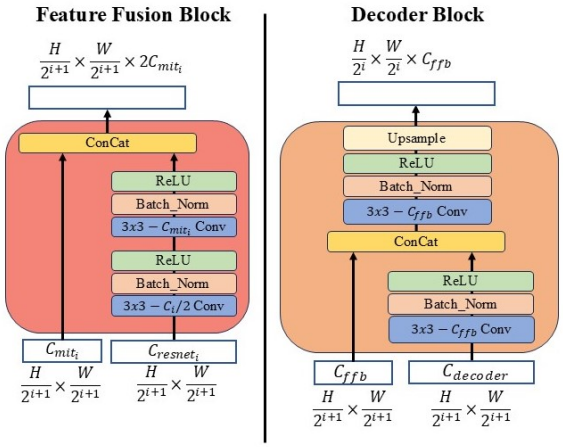


Figure 3. Feature Fusion Blocks and Decoder Blocks

block $i = \{1, 2, 3, 4\}$ can be shown as $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_{mit_i}$, where $C_{mit_i} \in \{64, 128, 320, 512\}$, respectively.

Feature Fusion Blocks: The feature maps from the two encoders at each scale differ in dimensions. Since the ResNet encoder produces feature maps with more channels than the MiT encoder, two convolutional layers are applied to the ResNet outputs, first halving the channels, then matching them to the MiT output, as shown in Fig. 3. The adjusted feature maps are then concatenated, resulting in a fused output of size $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times 2C_{mit_i}$.

Decoder Block: As shown in Fig. 3, each Decoder Block receives two inputs: one from the previous Decoder Block and one from the corresponding Feature Fusion Block. Before concatenation, the channel count of the input from the previous Decoder Block is reduced to match that of the Feature Fusion Block. After concatenation, a convolutional layer is applied, followed by spatial upsampling to prepare for the next Decoder Block. Finally, the output of the last Decoder Block is brought to the input resolution in the height head. Since height estimation is a regression problem, the Mean Squared Error (MSE) loss function is the preferred choice.

4.2. FusedSeg-HE

FusedSeg-HE aims to improve accuracy by incorporating a segmentation head into the Fused-HE method, as illustrated in Fig. 4. In height estimation, errors often occur at object boundaries, primarily because models struggle to correctly identify object locations and assign height values to the appropriate pixels. Although the predicted height values may be accurate, misalignment at object edges can significantly impact the error. To mitigate this issue, the segmentation head helps the model distinguish object pixels, ensuring height values are assigned to the correct locations. The ground truth for segmentation can be directly derived from nDSM data, where terrain height values are eliminated, leaving only object heights. This enables the creation

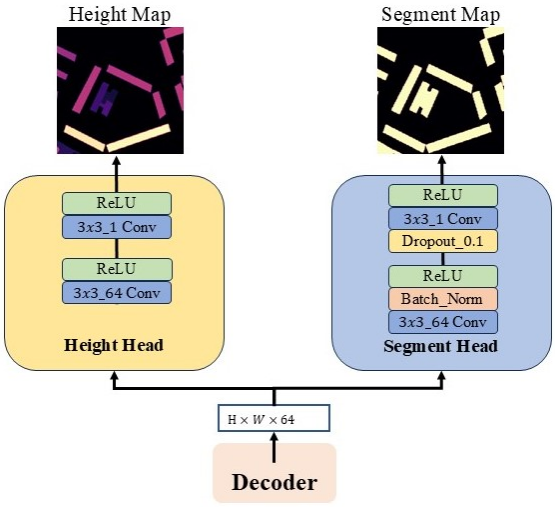


Figure 4. An illustration of the Segment Head and the Height Head of the FusedSeg-HE model.

of a binary segmentation mask by classifying terrain areas with a value of 0 and objects with values greater than 0.

As shown in Fig. 4, the segmentation and height heads share the same features extracted by the network. The height head uses the MSE loss function, while the segmentation head employs the Binary Cross Entropy (BCE) loss function. The total loss is computed as:

$$Loss_{total} = MSE(y_i, \hat{y}_i) + \lambda \cdot BCE(s_i, \hat{s}_i) \quad (1)$$

where y_i and $\hat{y}_i$ denote the target and predicted height maps in the i-th batch, respectively, while s_i and $\hat{s}_i$ represent the target and predicted segmentation binary masks in the same batch. The parameter λ controls the influence of the segmentation loss on the height prediction.

5. Results and Discussion

In this section, we compare the implemented following Fused-HE and FusedSeg-HE methods with three baseline models: a CNN-HE with a ResNet encoder, a ViT-HE based on the DPT architecture, and a MiT-HE using the SegFormer encoder. The baseline models use a single state-of-the-art encoder. For the sake of comparison, we use these encoders in the fusion process to demonstrate whether the fused encoder outperforms each individual encoder. Additionally, the Depth Anything model is fine-tuned for comparison. Tab. 1 provides details on all the trained models.

5.1. Experimental Setup

The Fused-HE models are trained on the DFC2019 and DFC2023 datasets using a learning rate of 1×10^{-5} and the Adam optimizer for 70 epochs. Batch sizes are set to 4 for DFC2019 and 1 for DFC2023. The FusedSeg-HE model is trained on the DFC2023 dataset with a learning rate of 5×10^{-5} using the Adam optimizer for 70 epochs with a batch size of 1.

Table 1. Encoder types of trained models and details of head units.

Models	Encoder Type	Encoder	Head	Params
CNN-HE	Conv	ResNet-101 [18]	Height	144M
ViT-HE	Standard ViT	DPT-Base [11]	Height	119M
MiT-HE	Hierarchi. ViT	MiT-B4 [43]	Height	101M
Fused-HE-RV	Conv+Standard ViT	ResNet-101+DPT-Base	Height	275M
Fused-HE-RM	Conv+Hierarchi. ViT	ResNet-101+MiT-B4	Height	181M
FusedSeg-HE-RM	Conv+Hierarchi. ViT	ResNet-101+MiT-B4	Height+Segment.	181M

5.2. Quantitative Evaluation

We evaluate model performance using multiple metrics, including Root Mean Square Error (RMSE), logarithmic RMSE (RMSE_{log}), delta (δ) threshold accuracy, logarithmic error ($\log_{10}$), Scale-Invariant logarithmic error (SI-log) and Intersection over Union (IoU). In addition, we use masked RMSE (RMSE_{mask}), which calculates the error only for pixels where the target height values exceed two meters. This helps eliminate the influence of background pixels, which have zero error.

Tab. 2 and Tab. 3 compare the performance of our fusion-based models against the baselines on the DFC2023 and DFC2019 datasets. As shown in Tab. 2, the Fused-HE Model, outperforms the baseline models and Depth Anything (vit-b) training from scratch. Despite being pre-trained on diverse depth-related datasets, including those related to drone imagery, the Depth Anything model achieved results comparable to those of our proposed Fused-HE model in fine-tuning. Notably, Fused-HE, although based on general-purpose ImageNet-pretrained encoders (ResNet and SegFormer MiT), outperformed Depth Anything in several metrics. Furthermore, our FusedSeg-HE model consistently surpassed both models across all evaluation metrics, demonstrating its superior generalization and effectiveness for satellite-based height estimation.

Tab. 3 shows results on the high-resolution DFC2019 dataset, containing imagery from two cities. MiT-HE again outperforms other baselines when trained from scratch, while fine-tuned ViT-HE achieves the best accuracy due to prior training on large aerial datasets. This highlights that standard ViTs need more data to generalize, whereas hierarchical ViTs like MiT are more sample-efficient. Our fusion method consistently achieves the best results across all metrics and training setups.

Tab. 4 analyzes the impact of λ on the performance of FusedSeg-HE. Higher λ values prioritize segmentation accuracy at the expense of height estimation, while lower values improve height estimation but reduce segmentation performance. The best balance is achieved at $\lambda = 0.005$, optimizing both segmentation and height estimation accuracy.

Our experiments highlight the importance of combining local and global feature extraction for height estimation from satellite imagery. Fused-HE method using ResNet-

Table 2. Results of different configurations of Fused-HE, FusedSeg-HE, and baseline methods on the DFC2023 dataset. Arrows indicate whether higher or lower values are preferable, with the best results in bold and the second-best underlined. Results of models trained from scratch are presented in the upper block, while results of models fine-tuned with pre-trained weights are shown in the lower block.

Models	Pre-Trained	$RMSE \downarrow$	$RMSE_{mask} \downarrow$	$RMSE \log \downarrow$	$IoU \uparrow$	$\log_{10} \downarrow$	$SI \log \downarrow$	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
CNN-HE	-	4.286	6.883	4.009	0.583	0.762	3.927	0.706	0.755	0.782
ViT-HE	-	4.812	7.799	4.120	0.539	0.798	4.061	0.696	0.742	0.769
Depth Anything [46]	-	4.611	7.459	4.233	0.553	0.811	4.001	0.671	0.718	0.748
MiT-HE	-	4.253	6.870	3.899	0.594	0.720	3.833	0.714	0.760	0.785
Fused-HE-RV	-	4.792	7.611	3.984	0.546	0.732	3.941	0.709	0.758	0.785
Fused-HE-RM	-	**4.206**	**6.852**	**3.852**	**0.598**	**0.716**	**3.779**	**0.719**	**0.765**	**0.789**
CNN-HE	✓	3.854	6.163	3.455	0.645	0.576	3.413	0.767	0.811	0.833
ViT-HE	✓	3.801	6.128	3.220	0.670	0.503	3.195	0.787	0.832	0.854
MiT-HE	✓	3.490	5.876	3.115	0.689	0.488	3.121	0.792	0.835	0.857
Depth Anything [46]	✓	3.407	5.508	2.846	0.715	0.441	2.846	0.802	0.846	0.867
Fused-HE-RV	ResNet-101+DPT-Base	3.730	6.036	3.334	0.659	0.529	3.305	0.774	0.826	0.855
Fused-HE-RV	DPT-Base	3.835	6.021	3.221	0.670	0.499	3.195	0.790	0.835	0.858
Fused-HE-RM	ResNet-101+MiT-B4	3.384	5.522	3.029	0.703	0.463	2.998	0.803	0.843	0.862
Fused-HE-RM	MiT-B4	3.379	5.524	2.932	0.706	0.431	2.911	0.810	0.850	0.871
FusedSeg-HE-RM	MiT-B4	**3.346**	**5.451**	**2.632**	**0.739**	**0.355**	**2.620**	**0.836**	**0.878**	**0.897**

Table 3. Results of different configurations of Fused-HE and baseline methods on the DFC2019 dataset. Arrows indicate whether higher or lower values are preferable, with the best results in bold and the second-best underlined. Results of models trained from scratch are presented in the upper block, while results of models fine-tuned with pre-trained weights are shown in the lower block.

Models	Pre-Trained	$RMSE \downarrow$	$RMSE_{mask} \downarrow$	$RMSE \log \downarrow$	$IoU \uparrow$	$\log_{10} \downarrow$	$SI \log \downarrow$
CNN-HE	-	2.953	4.470	5.819	0.436	1.645	5.391
ViT-HE	-	2.651	4.311	5.552	0.477	1.675	5.204
MiT-HE	-	2.687	4.158	5.419	0.496	1.481	5.045
Fused-HE-RM	-	**2.569**	**4.097**	**5.179**	**0.511**	**1.358**	**4.870**
CNN-HE	✓	2.597	4.011	5.048	0.523	1.287	4.777
ViT-HE	✓	2.129	3.713	4.114	0.618	0.998	4.012
MiT-HE	✓	2.322	3.847	4.3777	0.585	1.010	4.221
Fused-HE-RV	DPT-Base	2.216	3.590	4.760	0.597	1.218	4.479
Fused-HE-RM	MiT-B4	**2.057**	**3.586**	**4.018**	**0.634**	**0.921**	**3.994**

Table 4. Results of FusedSeg-HE-RM for different λ values in balancing the loss on the DFC2023 dataset. Arrows indicate whether higher or lower values are preferable, with the best results in bold and the second-best underlined.

λ	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	$RMSE \downarrow$	$RMSE_{mask} \downarrow$	$IoU \uparrow$
1	0.802	0.862	0.892	3.943	6.725	0.735
0.5	0.803	0.865	0.895	3.762	6.391	0.738
0.05	0.824	0.876	0.896	3.653	5.919	**0.755**
0.01	0.831	0.875	0.895	3.442	5.721	0.735
0.005	**0.836**	**0.878**	**0.897**	**3.346**	**5.451**	0.739
0.001	0.821	0.863	0.883	3.431	5.527	0.717
0.0005	0.819	0.860	0.880	3.456	5.509	0.713

101 and MiT-B4 encoders outperforms individual encoders by leveraging CNNs for edge-aware features and ViTs for global context, while FusedSeg-HE further boosts accuracy through segmentation. Hierarchical vision transformers like MiT enhance efficiency with deeper architectures and fewer parameters. Overall, hierarchical fusion proves highly effective, achieving state-of-the-art accuracy and surpassing baseline models.

5.3. Qualitative Evaluation

Fig. 5 shows five example satellite images and the height estimation results of Fused-HE, FusedSeg-HE, and the baseline methods, all trained on the DFC2023 dataset. For one image, two zoomed-in areas are also shown. The results indicate that CNN-based methods capture sharp features, such as building edges, while ViTs produce smoother and more accurate height estimations. Fused-HE outperforms both by effectively integrating local and global features. Also, the segmentation head in FusedSeg-HE further improves height assignment, especially at object boundaries. More results are in the supplementary materials.

Fig. 6 shows the height profiles of the FusedSeg-HE model trained on the DFC2023 dataset for three selected regions, plotted and compared to the ground truth profiles. The results show that the model accurately identifies building locations and assigns height values to the correct pixels. However, it tends to produce smoother surfaces, making it challenging to capture abrupt height changes over short distances. Further 3D visualizations of FusedSeg-HE predictions are provided in the supplementary material.

Generalizability: Our goal in this work is to develop robust height estimation methods for satellite and aerial images that generalize well across different datasets. Fig. 7 presents height estimations from the FusedSeg-HE model, trained on nadir-view satellite images from the DFC2023 dataset, applied to images from various sensors and real-world scenarios. The results on pre- and post-disaster images from the WorldView-3 (WV3) satellite clearly capture changes in building heights, demonstrating the accuracy of the estimations. Similarly, in the post-flood image, the height differences between affected and unaffected areas are distinctly visible. Additionally, the model's performance on high-resolution aerial images, which may include off-nadir angles, is shown in the last three rows. The predicted height values are promising, further validating the model's robustness. Overall, these results indicate that our proposed method can be effectively applied to real-world tasks such as disaster management, urban planning, and en-

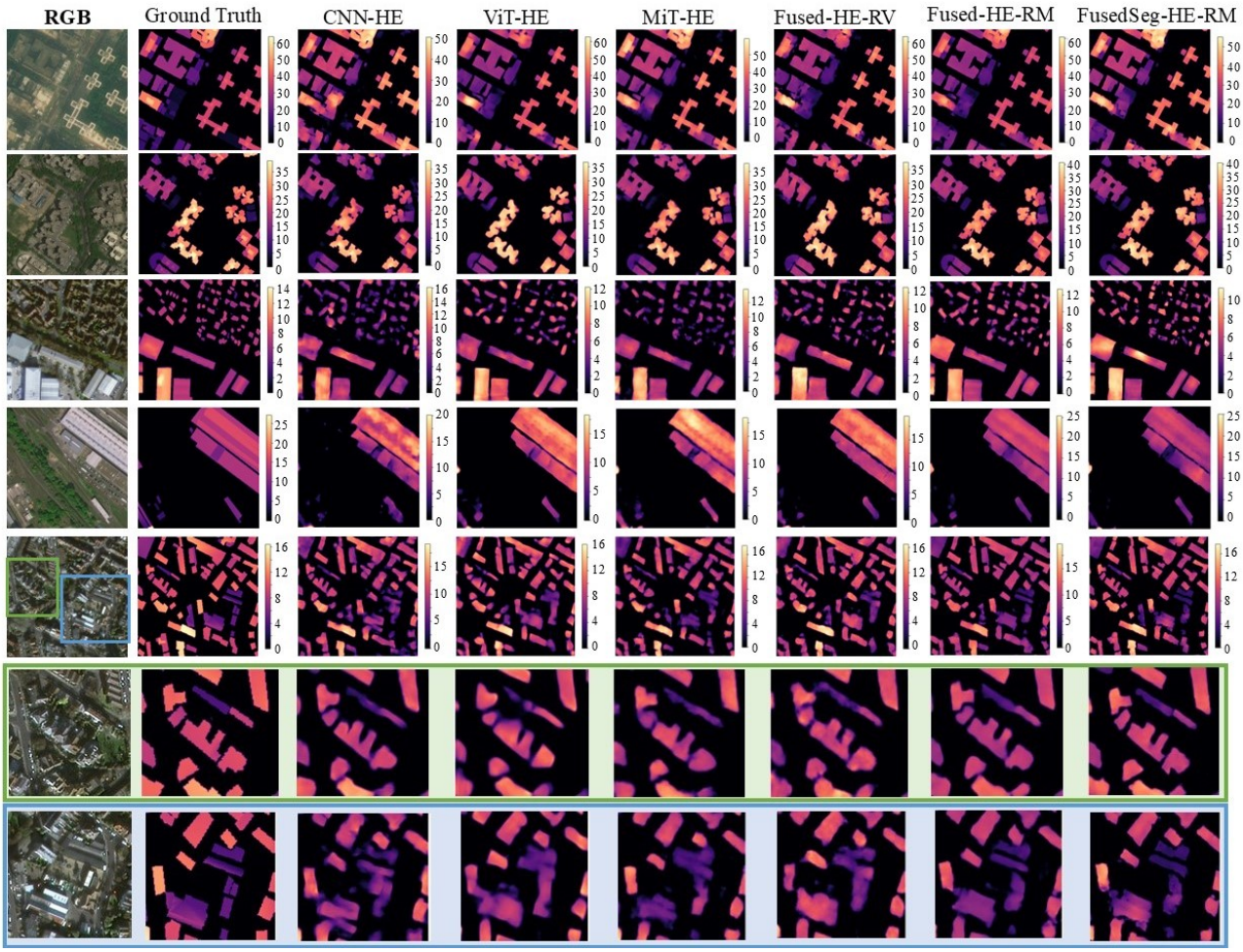


Figure 5. Results of Fused-HE, FusedSeg-HE, and baseline methods, all trained on DFC2023, shown on a test image from DFC2023 with two zoomed-in areas. The scale bars represent height values in meters.

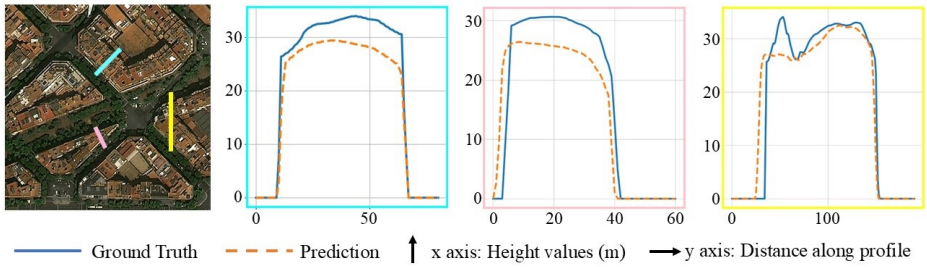


Figure 6. Height profile of selected buildings from a FusedSeg-HE-RM prediction and its ground truth on the DFC2023 test data.

vironmental monitoring.

Limitation: Fig. 8 presents the FusedSeg-HE results for an image containing buildings with different heights but similar visual characteristics. This challenge may arise from factors such as similar roof materials or shadows, which reduce feature discriminability. As shown in the predicted height values, the model struggles to differentiate between the buildings, resulting in similar height estimations for both.

6. Conclusions

This work presents Fused-HE and FusedSeg-HE, two methods designed for height estimation from satellite imagery by integrating convolutional and vision transformer encoders. Our experiments show that relying solely on CNNs or ViTs as encoders limits performance, whereas combining them

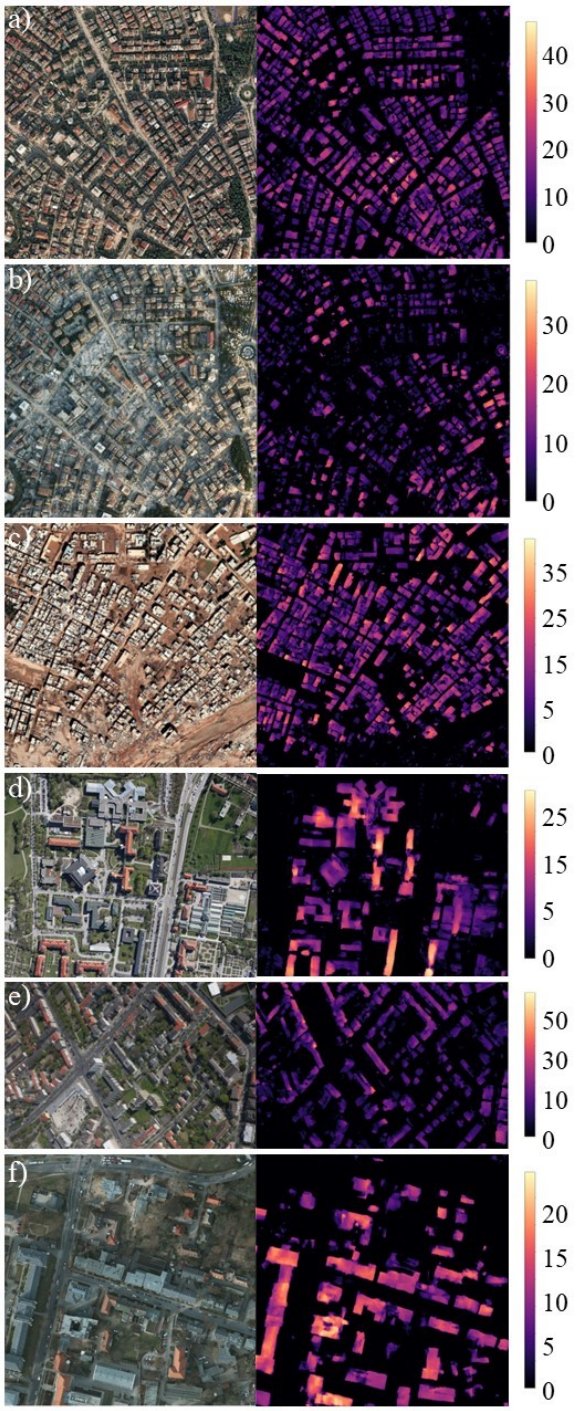


Figure 7. Results of the FusedSeg-HE-RM model on images from the WV3 satellite before and after the earthquake in Turkey [38] with 30 cm GSD (a, b), WV3 after flooding in Libya (c), nadir and oblique aerial images with 13 cm and 10 cm GSD from the SkyScapes and EAGLE datasets [2, 3] (d, e), and an ortho aerial image with 5 cm GSD from the Potsdam dataset [19] (f). The scale bars represent height values in meters.

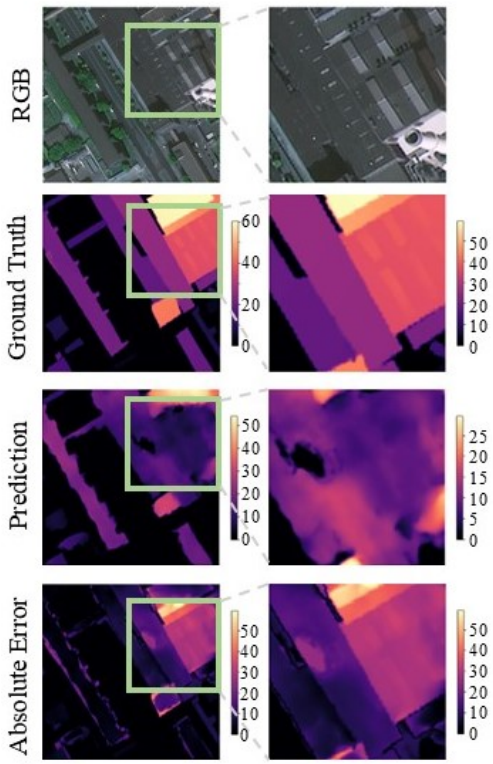


Figure 8. FusedSeg-HE-RM results on a satellite image from the DFC2023 test dataset, highlighting the model's difficulty in predicting accurate height values for objects with different heights but similar visual characteristics. The scale bars are in meters.

leverages the strengths of both, CNNs for capturing fine details and ViTs for global context, leading to significantly improved accuracy. The addition of a segmentation head in FusedSeg-HE further refines height predictions by improving pixel-level alignment, particularly at object boundaries.

Hierarchical vision transformers, like MiT, enhance efficiency with deeper architectures and fewer parameters while preserving multi-scale feature extraction. Our results show hierarchical fusion as highly effective, achieving state-of-the-art accuracy and outperforming baseline models across datasets. The models generalize well across real-world scenarios, including disaster areas, high-resolution aerial imagery, and diverse sensors like WV-3.

This study emphasizes the value of combining local and global feature extraction for height estimation and the synergy between segmentation and regression methods. Future work will focus on improving model efficiency and adaptability across diverse aerial and satellite datasets and imaging conditions.

Acknowledgement. The first author acknowledges the support and hospitality of the Remote Sensing Technology Institute of the German Aerospace Center (DLR), where a part of this research was conducted during his research visit.

References

[1] Hamid Ali Amirkolaee and Hamed Arefi. Height estimation from single aerial images using a deep convolutional encoder-decoder network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 149:50–66, 2019.

[2] Seyed Majid Azimi, Corentin Henry, Lars Sommer, Arne Schumann, and Eleonora Vig. Skyscapes fine-grained semantic understanding of aerial scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7393–7403, 2019.

[3] Seyed Majid Azimi, Reza Bahmanyar, Corentin Henry, and Franz Kurz. Eagle: Large-scale vehicle detection dataset in real-world scenarios using aerial imagery. In *2020 25th international conference on pattern recognition (ICPR)*, pages 6920–6927. IEEE, 2021.

[4] Jaewon Bae, Seungjun Moon, and Soohyun Im. Deep digging into the generalization of self-supervised monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 187–196, 2023.

[5] Sanaullah F. Bhat, Ismail 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.

[6] J. B. Campbell and R. H. Wynne. *Introduction to Remote Sensing*. Guilford Press, 2011.

[7] 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, 2019.

[8] Shengjie Chen, Yushan Shi, Zhiqiang Xiong, and Xiao Xiang Zhu. Adaptive bins for monocular height estimation from single remote sensing images. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 7015–7018. IEEE, 2023.

[9] Shengjie Chen, Yushan Shi, Zhiqiang 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.

[10] Xiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Hu Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *Advances in Neural Information Processing Systems*, pages 9355–9366, 2021.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint*, 2020.

[12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, 2014.

[13] Ravi Garg, Vijay Kumar B. G, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 740–756. Springer International Publishing, 2016.

[14] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[15] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[17] Vitor Guizilini, Ravi Hou, Jinkun Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. *arXiv preprint*, 2020.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[19] International Society for Photogrammetry and Remote Sensing (ISPRS). 2d semantic labeling contest - potsdam, 2023.

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

[21] Dongwon Kim, Woojin Ka, Pyo Ahn, Doyeon Joo, Sungjin Chun, and Junmo Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv preprint*, 2022.

[22] Maximilian Klingner, Jan A. Termöhlen, Jakub Mikolajczyk, and Tobias Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*, pages 582–600. Springer International Publishing, 2020.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolution neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

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

[26] Zhenyu Li, Xiaohong Wang, Xin Liu, and Jian Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint*, 2022.

[27] Cheng-Ju Liu, Victor A. Krylov, Philip Kane, Gerald Kavanagh, and Ramesh Dahyot. Im2elevation: Building height estimation from single-view aerial imagery. *Remote Sensing*, 12(17):2719, 2020.

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[29] Robert McCraith, Lukas Neumann, Andrew Zisserman, and Andrea Vedaldi. Monocular depth estimation with self-supervised instance adaptation. *arXiv preprint*, 2020.

[30] E. M. Mikhail, J. S. Bethel, and J. C. McGlone. *Introduction to Modern Photogrammetry*. John Wiley & Sons, 2001.

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

[32] Claudia Persello, Rasmus Hänsch, Giuseppe Vivone, Ke Chen, Zhiliang Yan, Da Tang, Hao Huang, Michael Schmitt, and Xiaoxiang Sun. 2023 ieee grss data fusion contest: Large-scale fine-grained building classification for semantic urban reconstruction. *IEEE Dataport*, 2022.

[33] 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.

[34] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021.

[35] Benoît Le Saux, Naoto Yokoya, Rasmus Hänsch, and Michael Brown. Data fusion contest 2019 (dfc2019). *IEEE Dataport*, 2019.

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, 2014.

[37] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[38] Maxar Technologies. Open data program: Turkey earthquake 2023. 2023.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[40] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[41] Yufeng Wang, Wenrui Ding, Ruiqian Zhang, and Hongguang Li. Boundary-aware multitask learning for remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:951–963, 2020.

[42] Huaxia Wu, Baoyuan Xiao, Noel Codella, Meng Liu, Xingchao Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021.

[43] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, pages 12077–12090, 2021.

[44] Siyuan Xing, Qiulei Dong, and Zhanyi Hu. Sce-net: Self- and cross-enhancement network for single-view height estimation and semantic segmentation. *Remote Sensing*, 14(9): 2252, 2022.

[45] Zhiqiang Xiong and Xiao Xiang Zhu. Knowledge transfer for label-efficient monocular height estimation. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 5377–5380. IEEE, 2022.

[46] 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.

[47] Chenyang Zhao, Yuhang Zhang, Matteo Poggi, Fabio Tosi, Xiaoxiao Guo, Zhixin Zhu, Gao Huang, Yike Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *2022 International Conference on 3D Vision (3DV)*, pages 668–678. IEEE, 2022.

[48] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017.