THE Benchmark: Transferable Representation Learning for Monocular Height Estimation

Zhitong Xiong, *Member, IEEE*, Wei Huang, Jingtao Hu, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*— Generating 3-D city models rapidly is crucial for many applications. Monocular height estimation (MHE) is one of the most efficient and timely ways to obtain large-scale geometric information. However, existing works focus primarily on training and testing models using unbiased datasets, which does not align well with real-world applications. Therefore, we propose a new benchmark dataset to study the transferability of height estimation models in a cross-dataset setting. To this end, we first design and construct a large-scale benchmark dataset for cross-dataset transfer learning on the height estimation task. This benchmark dataset includes a newly proposed large-scale synthetic dataset, a newly collected real-world dataset, and four existing datasets from different cities. Next, a new experimental protocol, *few-shot cross-dataset transfer*, is designed. Furthermore, in this article, we propose a scale-deformable convolution (SDC) module to enhance the window-based Transformer for handling the scale-variation problem in the height estimation task. Experimental results have demonstrated the effectiveness of the proposed methods in traditional and cross-dataset transfer settings. The datasets and codes are publicly available at https://mediatum.ub.tum.de/1662763 and https://thebenchmarkh.github.io/.**

***Index Terms*— Benchmark, cross-dataset transfer, remote sensing, synthetic data, transfer learning, Transformer.**

I. Introduction

MONOCULAR height estimation (MHE) [1] is of great importance to rapid 3-D city modeling and can give a basic insight into urbanization level. Geometric information from 3-D cities can be used for energy demand estimation, population estimation, damage forecasting, and so on [2].

Manuscript received 1 February 2023; revised 10 July 2023; accepted 2 August 2023. Date of publication 13 September 2023; date of current version 22 September 2023. This work was supported in part by the German Federal Ministry for Economic Affairs and Climate Action in the framework of the "National Center of Excellence ML4Earth" under Grant 50EE2201C, in part by the German Research Foundation (DFG GZ: ZH 498/18-1) under Project 519016653, in part by the German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV) based on a resolution of the German Bundestag (Acronym: *EKAPEx*) under Grant 67KI32002B, and in part by the German Federal Ministry of Education and Research (BMBF) in the framework of the International Future AI Laboratory "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001. *(Corresponding author: Xiao Xiang Zhu.)*

Zhitong Xiong, Wei Huang, and Xiao Xiang Zhu are with the Chair of Data Science in Earth Observation, Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: zhitong.xiong@tum.de; w2wei.huang@tum.de; xiaoxiang.zhu@tum.de).

Jingtao Hu is with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: jthu@mail.nwpu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3311764

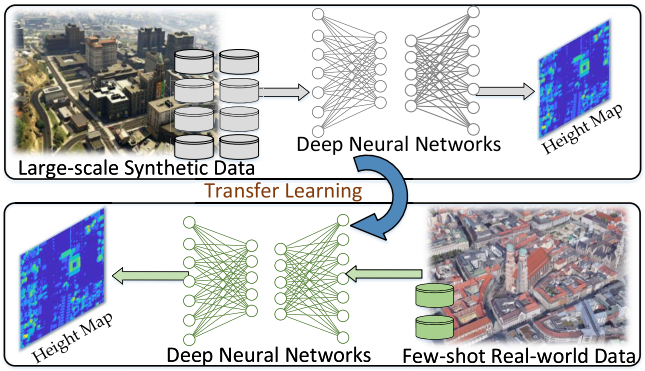


Fig. 1. Illustration of MHE in the cross-dataset transfer setting. We aim to transfer deep models from a large-scale synthetic dataset to different real-world datasets in a few-shot cross-dataset setting.

On the other hand, the generated height data can also contribute to many challenging follow-up research topics, such as urban planning [3], automatic piloting, and robot vision.

Airborne light detection and ranging (LiDAR) can actively acquire the digital surface model (DSM) data that contain accurate height information. However, LiDAR is cost-consuming, due to the airborne carrying platform and data storage device. Furthermore, its height information is not updated in a timely manner, limited by the hardware and complicated postprocessing process. Apart from LiDAR, incidental satellite images can also provide height data via the calculation of triangulation from satellite image pairs of consecutive views at different time intervals. Nevertheless, the conditions for obtaining such images are quite strict. In addition, both LiDAR and incidental images are difficult to handle complex city scenes in a real-time manner. In contrast, MHE can predict height using only a single aerial image and offers broad application potential in practice because of its fairly simple data acquisition requirements.

Benefiting from the powerful feature representation capacity, deep neural networks (DNNs) [4] have dominated most computer vision fields, including MHE. However, deep-learning-based MHE methods usually require a large amount of annotated data, which is difficult to obtain for several reasons.

First, it is very costly to obtain high-resolution remote sensing images and their corresponding ground-truth height values on a global scale. Due to different development levels and urban construction styles, different cities have their specific urban layouts, which leads to distinctive height distribution. This results in severe cross-city domain shifts that

are very common in real-world applications. One potential approach to mitigate these domain shifts is to gather data from multiple cities. However, it is prohibitively expensive or even impossible to collect high-quality data samples for the MHE task from diverse cities at a global scale.

Second, to ensure the performance of MHE models in real-world applications, it is imperative to construct datasets for MHE model training that encompass a wide range of imaging conditions. The existing methods usually train and test MHE models using unbiased datasets [5]. However, for real-world applications, providing the training data under diverse imaging conditions can be effective at improving the robustness and performance of deep networks. However, constructing datasets with different imaging conditions, such as different camera poses (heights and angles), camera resolutions, and viewing fields, is further expensive and difficult. Consequently, there is a lack of a high-resolution, highly accurate, and large-scale annotated height estimation benchmark dataset.

To address these aforementioned limitations, we resort to constructing a large-scale synthetic dataset that contains high-resolution images with accurate geometric information captured under different conditions. The presented benchmark dataset, termed **T**ransferable Monocular **H**eight **E**stimation (THE), can foster research on transferable representation learning for MHE, as illustrated in Fig. 1. In addition to the benchmark dataset, we design a new Transformer-based method to enhance the performance of MHE models in two cross-dataset experimental settings. To summarize, we make the following contributions.

1) Collecting and releasing two new datasets for MHE. One is a large-scale synthetic dataset termed Grand Theft Auto for Height estimation (GTAH), which is obtained from the game *Grand Theft Auto* [6], under different imaging conditions. The other dataset is a real-world one collected from the Actueel Hoogtebestand Nederland (AHN) project, which covers multiple cities in the Netherlands.
2) Constructing a new benchmark platform for transferable MHE. Specifically, one synthetic dataset and five real-world datasets are included to explore the feasibility of height knowledge transfer from synthetic to real scenes. We propose a few-shot cross-dataset transfer setting to evaluate deep models on datasets that were not seen during training.
3) To further enhance the model transferability in a cross-dataset experimental setting, we design a new scale-deformable convolution (SDC) module to enhance the Transformer networks with adaptive spatial context. The SDC module can learn to adjust the spatial context of representations adaptively across different datasets.

Extensive quantitative and qualitative results show that our framework outperforms the existing methods clearly, which indicates the effectiveness of the proposed methods. The remainder of this article is organized as follows. Section II reviews related works. Section IV introduces the details of the proposed method. In Section V, extensive experiments and analysis are presented to verify the proposed method comprehensively. Finally, this work is concluded in Section VI.

II. Related Work

Both monocular depth estimation (MDE) and MHE are geometry-related regression tasks; the former motivates the development of the latter to some extent. In this section, related works on MDE are first investigated, and then MHE is introduced.

A. Monocular Depth Estimation

Early works on MDE used hand-crafted visual features and probabilistic graphical models (PGMs) to encode depth-specific visual cues, including object size and texture density, based on a strong geometric assumption [7], [8], [9]. Recently, deep-learning-based methods have dominated this field because of their powerful feature representation capacity. There are roughly two types of deep-learning-based MDE, supervised methods [10], [11], [12], [13] and self-supervised (unsupervised) methods [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25].

In this work, we mainly introduce the supervised methods, which take a single image as input and generate a pixelwise depth prediction map, following the standard supervised-learning workflow with the need for manual-annotated depth labels. These methods achieve state-of-the-art performance by making breakthroughs in innovative architecture designs, effective incorporation of geometric and semantic constraints, and novel objective functions. Some enlightening works [10], [26], [27] have applied deep convolutional neural network (CNN) architectures to MDE, directly estimating depths from single monocular images in an end-to-end trainable manner, and achieving impressive performance. To model the semantic and geometric structure of objects within a scene, some work [28], [29], [30] has introduced semantic segmentation into MDEs as an auxiliary task, which can guide depth estimation at the object level.

Taking into account the imbalanced depth distribution that restricts model performance, Jiao et al. [12] presented an attention-based distance-related loss to concern more distant depth regions. Lee and Kim [31] combined multiple loss terms adaptively to train a monocular depth estimator from a constructed loss function space containing many loss terms. To balance the coverage speed of these losses, a loss-aware adaptive rebalancing algorithm was further designed in the course of training. The work most closely related to ours is [32], in which a robust training objective is designed to train deep-learning-based MDE models using multiple mixing datasets. For the first time, they propose to evaluate MDE models in a zero-shot cross-dataset transfer setting. More recently, vision Transformer-based deep models [33], [34] have also been proposed, to take advantage of the powerful representation learning ability of the Transformer backbone.

B. Monocular Height Estimation

Motivated by the success of deep-learning-based MDE, researchers have attempted to directly predict the height of objects, i.e., the DSM, within single aerial images from an overhead view [1], [35], [36], [37], [38], [39], [40], [41], [42],

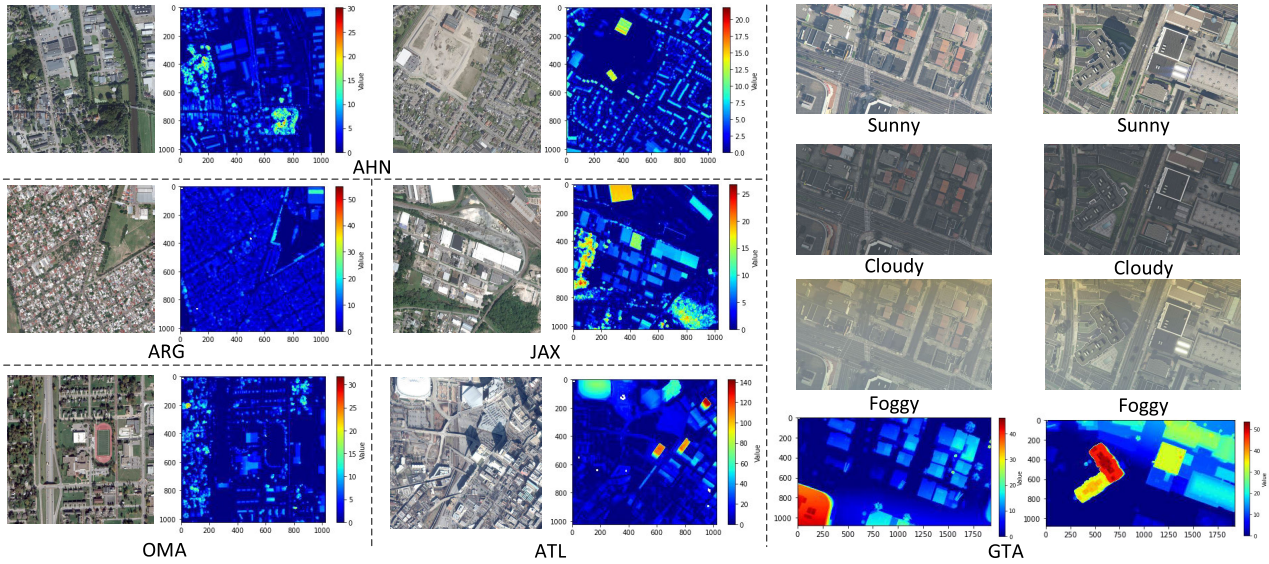


Fig. 2. Visualization of some samples from the synthetic GTAH dataset and real-world datasets. To simulate the complex real-world conditions in monocular aerial images, various imaging conditions, which may influence the performance of MHE, are taken into consideration, including different heights, weathers, and times of day. Moreover, it can be seen that the scale-variation problem between different datasets is very obvious.

[42], [43], [44], [45], [46]. Srivastava et al. [35] proposed a multitask CNN architecture for joint height estimation and semantic segmentation in monocular aerial images in 2017 for the first time. Mou and Zhu [1] published a concurrent work that proposed a fully convolutional–deconvolutional network for MHE and demonstrated its usefulness, for instance, the segmentation of buildings. Ghamisi and Yokoya [36] and Paoletti et al. [37] performed the image-to-image translation from monocular optical images to the corresponding depth maps within three cities, using the technique of generative adversarial network (GAN). Amini Amirkolaee and Arefi [38] presented a CNN-based method to identify collapsed buildings after an earthquake, based on preevent and postevent satellite images as well as airborne LiDAR data. Based on the CNN architecture, they further designed a postprocessing approach to merge multiple predicted height image patches into a seamless continuous height map [39]. Liu et al. [40] proposed a joint framework called height-embedding context reassembly network (HECR-Net) to simultaneously predict semantic labels and height maps from single aerial images, by distilling height-aware embeddings implicitly.

Leveraging the optical flow prediction technique, Chridtie et al. [41] developed an encoding strategy of the universal geocentric pose of objects within static monocular aerial images and trained a deep network to compute the dense representation; these attributes were exploited to rectify oblique images to dramatically improve the accuracy of height prediction of multiple images taken from different oblique viewpoints. Mahmud et al. [42] proposed a boundary-aware multitask deep-learning-based architecture for fast 3-D building modeling from single overhead images, by jointly learning a modified signed distance function, a dense height map, and scene semantics from building boundaries to model the buildings within the scenes. Madhuanand et al. [43] aimed to estimate depth from a single unmanned aerial vehicle (UAV) aerial image, by designing a self-supervised learning approach named self-supervised MDE that does not need any information other than images.

Although these works have contributed to the development of MHE in the past few years, most of them stayed within a particular comfort zone. There is an urgent need to study some more crucial issues that restrict the practical application of MHE in the open real world, such as the exploration in few-shot knowledge transfer in a cross-dataset setting, and scale-adaptive MHE model design. Aiming to address these problems, this article conducts the corresponding research and exploration.

III. Transferable Monocular Height Estimation

This section describes the proposed THE benchmark, which includes a synthetic GTAH dataset and five real-world datasets. The newly constructed GTAH and AHN datasets are described in detail, including their data source, dataset details, and statistical characteristics. Then to fairly evaluate the proposed two datasets in the MHE field, a comprehensive statistical analysis for datasets from five different domains is performed.

A. GTAH

In this section, the data source and dataset details of the synthetic GTAH dataset are introduced in detail.

1) Data Source: GTAH was collected from an electronic computer game called Grand Theft Auto V (GTA5) that was developed by Rockstar North and published by Rockstar Games [6]. The virtual world in GTA5 covers an area of 252 km^2, containing many scenes such as beach, stadium, mall, and store.

2) Dataset Details: GTAH contains a total of 85 881 pairs of synthetic monocular aerial images and their associated pixelwise height maps, with a resolution of 1920 × 1080. To simulate the complex real-world conditions in monocular

aerial images, various imaging conditions are taken into consideration, as follows.

1) *Height Distribution:* Most of the scenes in GTAH are located in areas with a rich variety of buildings for height diversity. Diverse height facilitates a comprehensive and fair assessment of the performance of MHE algorithms, for the situation where height estimation is valid for some heights but is poor when faced with a wide variety of heights.
2) *Camera Locations:* For the diversity of height and scene information, 1111 positions are selected along the roads in GTA5's city as the plane coordinates of the camera without respect to camera heights.
3) *Camera Angles:* Taking $[x, y, z]$ as the coordinate system, monocular aerial images of GTAH are captured from different viewpoints to simulate the diversity of camera pose.
4) *Camera Heights:* For the diversity of scene scales, monocular images are acquired at four camera heights of 300, 380, 460, and 540 m, to study the effects of the camera height in practice.
5) *Weather Types:* To evaluate the effectiveness and robustness of MHE methods in different weathers, GTAH contains three common weather conditions: "sunny," "foggy," and "cloudy."
6) *Shadows:* Shadow is an implicit visual cue influencing the performance of MDE models. Dijk and Croon [47] made an ablation study of shadows to demonstrate their effect on MDE. As a similar task focusing on pixelwise regression, it could be presumed that MHE may also be influenced by shadows. To enable a study of this kind, GTAH contains the monocular images with (w/) and without (w/o) shadows.
7) *Capturing Different Times of Day:* Different times of day have a significant impact on light intensity and direction, which further determines the shadow direction of buildings. For the fine study of capturing times of day and their subsequent effects, three times of day are considered in GTAH: 9:00 A.M., 15:00 P.M., and 18:00 P.M.

Some example images of GTAH are shown in Fig. 2, and some statistical results of GTAH are shown in Fig. 3.

B. AHN

The data source, collection, and properties of the real-world AHN dataset are introduced in this section.

1) Data Source: AHN was collected from the Actueel Hoogtebestand Nederland[1] project.

2) Dataset Details: AHN contains a total of 10 775 pairs of real monocular aerial images and their associated pixelwise height maps, with a resolution of 1024×1024. In the AHN dataset, images are selected to cover different scene types, including buildings, farms, forests, and water areas. The corresponding height maps are also carefully processed for MHE model training and evaluation. In addition, unlike other datasets, the AHN dataset covers multiple cities in the Netherlands, as shown in Fig. 2.

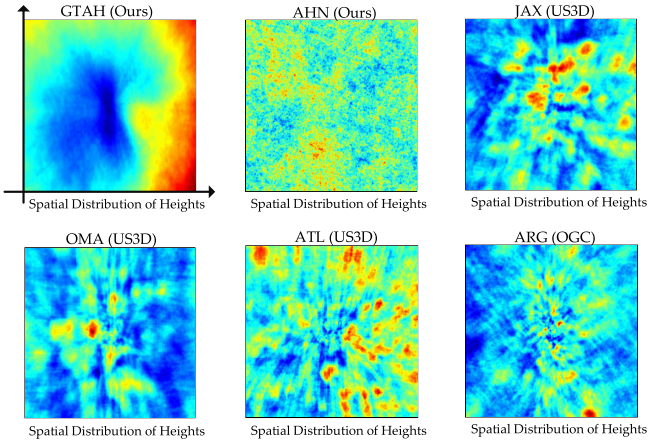


Fig. 3. Spatial distributions of heights of MHE datasets, including GTAH, AHN, JAX, OMA, ATL, and ARG. The x- and y-axes represent the width and height of the image, respectively. Darker colors represent larger height values. Different cities have clearly different spatial patterns.

3) How the Height Maps are Generated: GTAH is generated using the GTA game, where we manipulate the pose of cameras and adjust other rendering parameters to obtain a variety of RGB images along with their corresponding height maps. The height data in GTAH are synthetically generated using a game engine, ensuring high quality and precision. The AHN dataset is acquired from the AHN project, which uses airborne LiDAR technology. The height data in the AHN project undergo extensive quality control checks before being released, ensuring its reliability and accuracy. For the US3D datasets, the height data are derived from airborne LiDAR data obtained from the Homeland Security Infrastructure Program.[2] Specifically, the above ground level (AGL) height images are considered as the ground-truth height data.

C. Comparison With Other Existing MHE Datasets

To enrich the proposed THE benchmark dataset, we further take in the data of four cities from Urban Semantic 3-D [48], [49], [50], including Jacksonville (JAX), Omaha (OMA), Atlanta (ATL), and Argentina (ARG). The detailed comparisons of these six MHE datasets are presented in Table I. In Fig. 3, we visualize the spatial distributions of the height in different datasets. The ARG dataset is collected from the Overhead Geopose Challenge (OGC).[3] In addition, we have also analyzed the histogram of height distribution in six different datasets. In Fig. 4, we can see that the height distributions of all the six datasets obey long-tail distribution. It can be seen that the differences among these six datasets are clearly apparent. This also clearly indicates the domain shifts between different cities and the difficulty of cross-dataset transfer setting.

[1] https://www.ahn.nl/het-verhaal-van-ahn

[2] https://www.ahn.nl/hoe-werkt-het-inwinnen-van-hoogtegegevens

[3] https://www.nasa.gov/overhead-geopose-challenge

TABLE I
COMPARISON BETWEEN THE PROPOSED DATASETS AND OTHER EXISTING MHE DATASETS

Datasets	**GTAH (Ours)**	**AHN (Ours)**	JAX (US3D)	OMA (US3D)	ATL (US3D)	ARG (OGC)
Number of Images	**85,881**	**10,755**	1,098	1,796	704	2,325
Image Size	1920×1080	1024 × 1024	1024 × 1024	1024×1024	1024×1024	1024× 1024
Vertical/Oblique	Vertical & Oblique	Vertical	Oblique	Oblique	Oblique	Oblique
Weather Types	**Multiple**	Single	Single	Single	Single	Single
Daytimes	**Multiple**	Single	Single	Single	Single	Single
Height Scale	[0,439.2]	[0,195.8]	[0,186.5]	[0,194.1]	[0,123.4]	[0,92.7]
Shadows	**w/ & w/o**	w/	w/	w/	w/	w/
Real/Synthetic	Synthetic	Real	Real	Real	Real	Real
Number of Cities	Single	**Multiple**	Single	Single	Single	Single

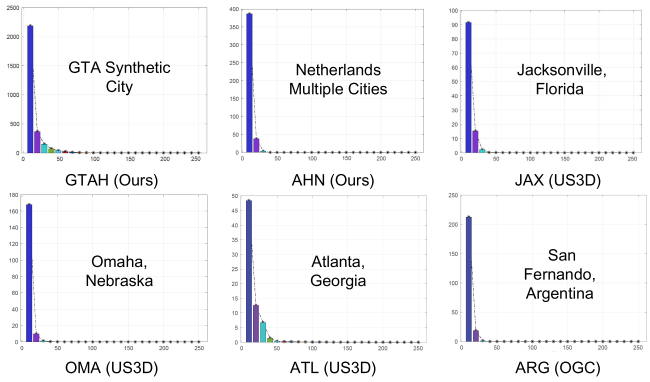


Fig. 4. Height distributions of six different MHE datasets, including GTAH, AHN, JAX, OMA, ATL, and ARG.

IV. METHODOLOGY

In this section, we will introduce the proposed transformer-based frameworks for MHE. The whole network architecture is shown in Fig. 5. Specifically, we will first introduce some existing vision transformers and their limitations for the MHE task. Then, an adaptive-structure convolution module is introduced to improve the performance and transferability of MHE. Finally, based on the constructed synthetic dataset, a few-shot cross-dataset transfer learning method is designed.

A. Vision Transformer for Height Prediction

Compared with CNN-based deep architectures, Transformers enable relationship modeling between input tokens and can capture relative height information through the self-attention mechanism. To predict the pixelwise height value from monocular images accurately, it is beneficial to make use of the relative height relationships between neighboring pixels. For the MHE task, this advantage makes Transformer-based architectures more effective in improving both the performance and the transferability of deep models.

B. Scale-Deformable Convolution for Few-Shot Cross-Dataset Transfer

For real-world applications, few-shot cross-dataset performance is a more faithful evaluation metric than training and testing on datasets with the same biases. Compared with the MDE task, the cross-dataset evaluation for MHE is more challenging. The reason is that remote sensing imagery captured at different heights will be greatly different due to the change in resolutions. However, the height values of objects on the ground should not change with different camera poses. This inconsistency makes MHE an extremely challenging task, especially in cross-dataset evaluation settings.

The Swin Transformer (Swin-T) can greatly reduce the computational complexity by computing self-attention maps within local windows. The window size is an important hyperparameter for the window-based Transformer models. However, for images with significantly different resolutions, aggregating context information in fixed-size windows has an obvious limitation: the spatial context for objects with different scales will be inconsistent. This makes the standard window-based Transformer less effective for handling the scale-variation problem. Consequently, the severe scale-inconsistent problem in remotely sensed images (as shown in Fig. 2) makes the fixed window size for image partitioning ineffective.

Considering this limitation, in this work, we propose an SDC module to adjust the spatial context of each pixel for the Transformer model in a learnable way. We achieve this goal by designing a deformable convolution operation with learnable dilation rates to adjust the context in a structured way. Given the input feature map $\mathbf{X} \in \mathbb{R}^{c\times h\times w}$ and kernel weight $\boldsymbol{w} \in \mathbb{R}^{c_o\times c\times k\times k}$, the standard convolution can be formulated as

$$\boldsymbol{V}_{\boldsymbol{p}_0} = \sum_{\mathbf{p}_n\in\Omega} \boldsymbol{w}(\boldsymbol{p}_n)\cdot \boldsymbol{x}(\boldsymbol{p}) \tag{1}$$

where $\boldsymbol{V}_{\boldsymbol{p}_0} \in \mathbb{R}^c$ denotes the output features at pixel $\boldsymbol{p}_0$. Indexes of the 2-D spatial offsets for the convolution operation are denoted by Ω. For a point $\boldsymbol{p}_0$ in the output feature map, the coordinates used for convolution computation are $\mathbf{p} = \mathbf{p}_0 + \boldsymbol{p}_n$.

To adjust the spatial context, the deformable convolution [51] was proposed to learn additional offsets to get a more flexible receptive field for each pixel. Although deformable convolution can learn adaptive context by offsets, we argue that merely using the offsets is still inefficient to adjust the receptive field with significant scale variation. For the traditional deformable convolution, usually, multiple deformable convolution layers are required to gradually adjust the context by the learned offsets. While using a learnable multiplier for the receptive field would be more effective, especially in the few-shot transfer settings. Thus, in this work,

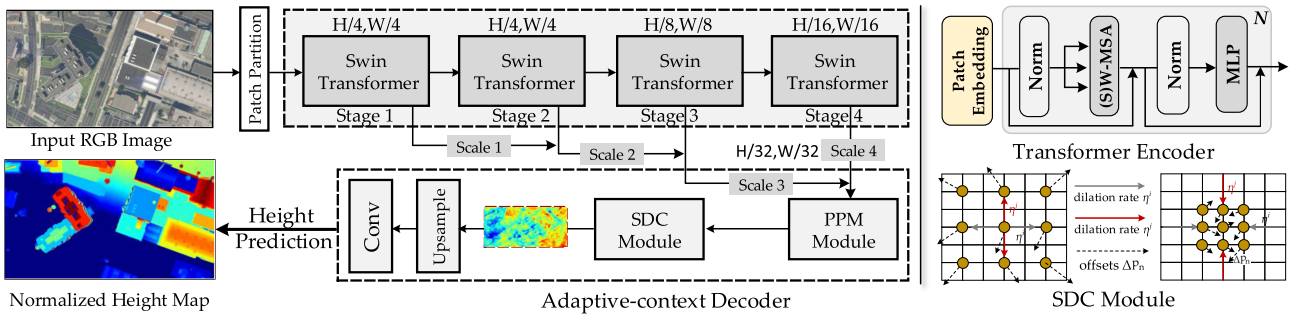


Fig. 5. Whole pipeline of the proposed "SwinUper + SDC." First, the height-related visual semantic feature is extracted by Swin-T. Then, adaptive scale modeling for covering a dynamic receptive field is achieved by the SDC module. Finally, the model is trained for height estimation using different types of loss functions for both the classical and cross-shot transfer settings.

we further extend the deformable convolution with learnable dilation rates. Based on this idea, the coordinates of point p becomes

$$\begin{aligned} \boldsymbol{p}^i &= \boldsymbol{p}_0^i + \eta^i \boldsymbol{p}_n^i + \Delta \boldsymbol{p}_n^i \\ \boldsymbol{p}^j &= \boldsymbol{p}_0^j + \eta^j \boldsymbol{p}_n^j + \Delta \boldsymbol{p}_n^j \end{aligned} \tag{2}$$

where $\eta = \{\eta^i, \eta^j\}$ are the learnable dilation rates, which can be used to control the receptive field for each pixel in a structured manner. The 2-D offsets of the deformable convolution are expressed by $\Delta \boldsymbol{p}_n = \{\Delta \boldsymbol{p}_n^i, \Delta \boldsymbol{p}_n^j\}$.

In practice, the dilation rates η and offsets $\Delta \boldsymbol{p}_n$ are typically fractional. To enable their end-to-end optimization, we adopt differentiable bilinear sampling to perform SDC, which can be defined as

$$\begin{aligned} \boldsymbol{V}_{\boldsymbol{p}_0}^c = \sum_u^H \sum_v^W \boldsymbol{x}^c(u, v) \max\left(0, 1 - |\boldsymbol{p}^i - v|\right) \\ \max\left(0, 1 - |\boldsymbol{p}^j - u|\right) \end{aligned} \tag{3}$$

where $\boldsymbol{p}_0 \in \{1, 2, \ldots, HW\}$ is the index of output feature maps, and $c \in [1 \ldots C]$ is the index of feature channels. For the sake of simplicity, the coordinate $\boldsymbol{p}_0$ will be omitted in the following formulas. The coordinates of input feature maps are denoted by u, v. Note that coordinates $(\boldsymbol{p}^i, \boldsymbol{p}^j)$ and (u, v) are normalized in the range of $[-1, 1]$. During backward propagation, we need to compute the partial derivatives with respect to $\boldsymbol{x}^c(u, v)$, $\boldsymbol{p}^i$, $\boldsymbol{p}^j$, $\Delta \boldsymbol{p}_n^i$, $\boldsymbol{p}_n^j$, η^i, and η^j. Based on (3), the partial derivatives for $\boldsymbol{x}^c(u, v)$ can be easily obtained by

$$\begin{aligned} \frac{\partial \boldsymbol{V}^c}{\partial \boldsymbol{x}^c(u, v)} = \sum_u^H \sum_v^W \max\left(0, 1 - |\boldsymbol{p}^i - v|\right) \\ \max\left(0, 1 - |\boldsymbol{p}^j - u|\right). \end{aligned} \tag{4}$$

Next, the partial derivatives of $\boldsymbol{p}^i$ can be computed by

$$\frac{\partial \boldsymbol{V}^c}{\partial \boldsymbol{p}^i} = \sum_{u,v}^{H,W} \boldsymbol{x}^c(u, v) \max\left(0, 1 - |\boldsymbol{p}^j - u|\right) g\left(v, \boldsymbol{p}^i\right) \tag{5}$$

where $g(v, \boldsymbol{p}^j)$ is a piecewise function that can be formulated as

$$g\left(v, \boldsymbol{p}^i\right) = \begin{cases} 0, & \text{if } |v - \boldsymbol{p}_i| \geq 1 \\ 1, & \text{if } v \geq \boldsymbol{p}_i \\ -1, & \text{if } v < \boldsymbol{p}_i. \end{cases} \tag{6}$$

The partial derivative $(\partial \boldsymbol{V}_k^c / \partial \boldsymbol{p}^j)$ is similar to that of $\boldsymbol{p}^i$. Furthermore, the partial derivative of η^i can be obtained by applying the chain rule

$$\begin{aligned} \frac{\partial \boldsymbol{V}^c}{\partial \eta^i} &= \frac{\partial \boldsymbol{V}^c}{\partial \boldsymbol{p}^i} \frac{\partial \boldsymbol{p}^i}{\partial \eta^i} \\ \frac{\partial \boldsymbol{p}^i}{\partial \eta^i} &= \boldsymbol{p}_n^i. \end{aligned} \tag{7}$$

The computation of partial derivatives $(\partial \boldsymbol{V}^c / \partial \eta^j)$ is similar to that of $(\partial \boldsymbol{V}^c / \partial \eta^i)$. Finally, we follow the same formulas described in deformable convolutional networks (DCN) [51] to compute the partial derivatives $(\partial \boldsymbol{V}^c / \partial \Delta \boldsymbol{p}_n)$.

C. Scale-Invariant Training Loss

Different from the MDE task, the MHE datasets may contain images captured at diverse camera poses, e.g., different camera heights. Due to the fact that the range of object heights may vary greatly for different camera poses, as shown in Fig. 2, it will be difficult to learn consistent deep representations for the MHE model across different camera poses. In this situation, better deep representations can be learned by training MHE models with consideration to relative height relationships. Thus, during the training stage, the loss functions consist of two components. The first loss term is the regular height map regression loss, which is defined with a pixelwise MSE loss. The second loss term is the scale-invariant training loss between different pixel pairs. Let $\boldsymbol{y}_h$ denote the ground-truth height map, and $\hat{\boldsymbol{y}}_h$ be the predicted height map. Then the final loss function can be defined as

$$\mathcal{L} = \mathcal{L}_{\text{mse}}\left(\hat{\boldsymbol{y}}_h, \boldsymbol{y}_h\right) + \mathcal{L}_{\text{rh}}\left(\hat{\boldsymbol{y}}_{hi}, \boldsymbol{y}_{hj}\right) \tag{8}$$

where $\mathcal{L}_{\text{mse}}$ denotes the MSE loss function. $\mathcal{L}_{\text{rh}}$ represents the relative height consistency loss. In the following part, we introduce three different implementations of the scale-invariant loss term including $\mathcal{L}_{\text{si}}$, $\mathcal{L}_r$, and $\mathcal{L}_{\text{msg}}$. These loss functions are initially proposed for the MHE task; in this work, we adapt them for the MHE task for performance comparison.

To handle the varying scale problem in training depth estimation models, Eigen et al. [52] designed a scale-invariant loss

$$\mathcal{L}_{\text{si}}\left(\hat{\boldsymbol{y}}_h, \boldsymbol{y}_h\right) = \frac{1}{n} \sum_i R_i^2 - \frac{1}{n^2} \left(\sum_i R_i \right)^2 \tag{9}$$

TABLE II

NUMBER OF TRAINING SAMPLES FOR FEW-SHOT CROSS-DATASET TRANSFER

Datasets	Training(#images)		Validation (#images)	Test (#images)
	1%	5%	—	—
ARG	16	80	200	725
ATL	4	20	200	304
JAX	6	30	200	498
OMA	12	60	200	596
AHN	50	250	200	3871

TABLE III

COMPARISON RESULTS OF DIFFERENT METHODS ON THE GTAH DATASET. THE BEST AND SECOND-BEST RESULTS ARE IN BLUE AND GREEN

Method	**Height Estimation Metrics**			
	MAE	**RMSE**	**SI-RMSE**	**MSGE**
U-Net (ResNet34) [55]	4.860	6.731	39.511	3.357
Adabins (ResNet50) [55]	3.651	5.552	28.125	2.957
DenseViT [55]	3.253	4.898	23.327	2.753
SwinUper [56]	2.946	4.628	21.169	2.358
SwinUper + $\mathcal{L}_{msg}$ [32]	2.943	4.614	20.754	2.277
SwinUper + $\mathcal{L}_{si}$ [52]	2.958	4.624	20.855	2.392
SwinUper + $\mathcal{L}_{r}$ [53]	2.995	4.683	21.302	2.371
SwinUper + SDC (Ours)	2.928	4.562	19.990	2.336

where R_i is the difference between the prediction and ground truth at pixel i, and $R_i = \boldsymbol{y}_{hi} - \hat{\boldsymbol{y}_{hi}}$. Note that for MHE, we use the original height value instead of the log space, as a large portion of the height values are zero.

Chen et al. [53] proposed to analyze the ordinal relationships and enforce the model to learn relative depth

$$\mathcal{L}_r = \begin{cases} \log\left(1+\exp\left(-\hat{\boldsymbol{y}_{i_k}} + \hat{\boldsymbol{y}_{j_k}}\right)\right), & r_k = +1 \\ \log\left(1+\exp\left(\hat{\boldsymbol{y}_{i_k}} - \hat{\boldsymbol{y}_{j_k}}\right)\right), & r_k = -1 \\ \left(\hat{\boldsymbol{y}_{i_k}} - \hat{\boldsymbol{y}_{j_k}}\right)^2, & r_k = 0. \end{cases} \tag{10}$$

This relative constraint loss $\mathcal{L}_r$ encourages the predicted depth map to agree with the ground-truth ordinal relationships.

Ranftl et al. [32] proposed to use gradient matching loss [54] to train the depth estimation models in the zero-shot cross-dataset transfer setting. In this work, the multiscale gradient matching loss is adapted to the MHE task by

$$\mathcal{L}_{\text{msg}}\left(\hat{\boldsymbol{y}_h}, \boldsymbol{y}_h\right) = \frac{1}{M}\sum_{k=1}^{K}\sum_{i=1}^{M}\left(|\nabla_x R_i^k| + |\nabla_y R_i^k|\right) \tag{11}$$

where R_i^k is the difference between the predicted height map and the ground-truth height map at scale k. The number of pixels in a predicted height map is denoted by M. In this work, four different scales [1, (1/2), (1/4), (1/8)] are used. We also combine the gradient matching loss $\mathcal{L}_{\text{msg}}$ with a standard MSE loss to form the final training loss.

V. EXPERIMENTS

This section begins by introducing the experimental settings. Then, the evaluation metrics are defined in brief. Finally, the few-shot synthetic-to-real transfer experiments are conducted, based on the existing deep semantic models and the proposed method.

TABLE IV

QUANTITATIVE RESULTS OF SWINUPER + SDC (GTAH) FOR ZERO-SHOT CROSS-DATASET TRANSFER FROM GTAH TO REAL CITIES

Datasets	MAE	SI-RMSE	RMSE	MSGE
ARG	5.42	39.61	9.38	7.54
ATL	17.6	156.7	14.57	13.05
JAX	9.43	71.58	11.92	7.13
OMA	4.73	51.76	8.35	5.41
AHN	2.66	22.18	6.24	5.86

A. Experimental Settings

In this section, a series of experiments are set up to evaluate the transferability of MHE models comprehensively.

1) *Benchmark Experiments on the GTAH Dataset:* Extensive experiments are conducted on GTAH to compare the effectiveness of different existing deep architectures, relative height loss functions, and the proposed method in this work.
2) *Few-Shot Cross-Dataset Transfer Experiments:* Experiments under the few-shot cross-dataset setting are performed to examine the transfer performance from the GTAH to real datasets when only a few annotated target samples are available. In addition, to better understand the effect of the proposed SDC module, we visualize and analyze the module's scale-adaptive ability for obtaining the adaptive spatial context.
3) *Pretraining Comparison Experiments:* Finally, to verify the superiority of the GTAH to ImageNet for pretraining the MHE models, their training losses and visualization of model weight distribution are provided intuitively.

B. Implementation Details

All the deep models are implemented in PyTorch. For the GTAH dataset, 100 epochs are used to train the deep models used for transfer learning in this work. For the CNN-based U-Net model with the ResNet-34 backbone, we use the code[4] from [41]. Adam [56] is used for optimizing the model with an initial learning rate of $1e-4$. For the SwinUper backbone, the tiny version of the Swin-T is used and the UperNet [57] is used for the decoder. The optimizer AdamW [58] is used with an initial learning rate of $6e-5$ for training all the Transformer-based deep models. Adabins [59] and DenseViT [33] are the state-of-the-art MDE models selected for performance comparison on the proposed GTAH dataset. The detailed hyperparameters for model training can be found in the publicly available code.[5]

In the few-shot cross-dataset transfer experiments, we randomly select 1% and 5% of the training data for each of the five real-world datasets (cities), as presented in Table II. Then 15 epochs are used to fine-tune the deep models initialized

[4]https://github.com/pubgeo/monocular-geocentric-pose
[5]https://github.com/EarthNets/3D-Understanding

TABLE V

EXPERIMENTAL RESULTS ON THE AHN DATASET IN THE FEW-SHOT CROSS-DATASET TRANSFER SETTING. THE RESULTS OF USING 1% AND 5% TRAINING DATA ARE REPORTED. THE BEST RESULTS ARE IN BLUE, AND THE SECOND-BEST ONES ARE IN GREEN

Methods	Height Estimation Metrics (1% Training)				Height Estimation Metrics (5% Training)			
	MAE	SI-RMSE	RMSE	MSGE	MAE	SI-RMSE	RMSE	MSGE
U-Net (ImageNet) [55]	2.348	8.739	5.759	1.620	2.313	9.988	5.600	1.564
U-Net (GTAH) [55]	2.289	8.334	5.673	1.536	2.256	9.716	5.632	1.520
Adabins (ImageNet) [60]	2.316	8.707	5.677	1.583	2.275	9.531	5.541	1.537
Adabins (GTAH) [60]	2.273	8.345	5.568	1.530	2.136	9.232	5.461	1.499
DenseViT (ImageNet) [33]	2.524	8.873	5.907	1.627	2.415	9.990	5.769	1.617
DenseViT (GTAH) [33]	2.439	8.633	5.986	1.616	2.314	9.217	5.640	1.536
SwinUper (ImageNet) [56]	2.394	9.002	5.694	1.724	2.104	9.297	5.585	1.435
SwinUper (GTAH) [56]	2.156	8.303	5.564	1.455	2.107	9.021	5.569	1.426
SwinUper+$\mathcal{L}_{msg}$ (GTAH) [32]	2.160	8.343	5.584	1.467	2.112	9.040	5.569	1.430
SwinUper+SDC (GTAH) (Ours)	2.117	8.285	5.530	1.441	2.001	8.981	5.439	1.417

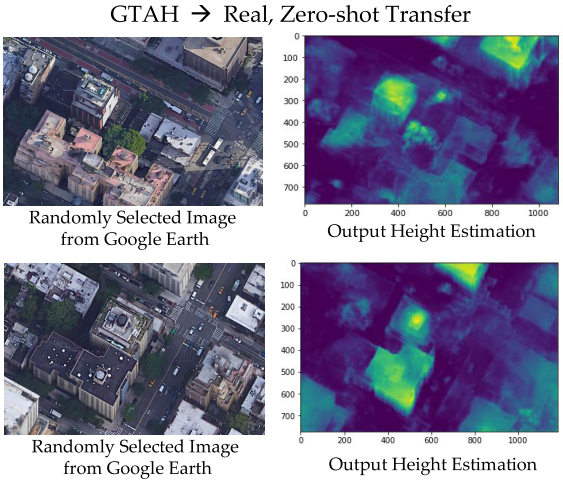


Fig. 6. Visualization of some height estimation samples on a few RGB images randomly selected from Google Earth. The MHE model is pretrained on the GTAH dataset, and then directly used to predict the normalized height maps of these images without fine-tuning.

with ImageNet or GTAH pretrained parameters. Finally, for each dataset, the full test set is used for evaluating the fine-tuned models. Other details of the proposed method can be found in the publicly available code.

C. Evaluation Metrics

Different from the MDE task, the area with height value 0 accounts for a large percentage of the image. To evaluate the effectiveness of the proposed methods, we propose to use four metrics on each dataset for performance evaluation: mean absolute error (MAE), root mean squared error (RMSE), scale-invariant RMSE (SI-RMSE) [52], and multiscale gradient error (MSG). Of these, MAE and RMSE are measurements that are widely used for the evaluation of regression tasks. MAE, defined as $\text{MAE} = 1/n * \sum |y_i - \hat{y}_i|$, is used to measure the mean absolute difference between the predicted height values and the ground-truth values in the whole dataset. RMSE, defined as $\text{RMSE} = (\Sigma(y_i - \hat{y}_i)^2/n)^{1/2}$, is more sensitive to large height values. We also propose to use SI-RMSE and MSGE to measure the relative relationships in the predicted height maps. SI-RMSE is defined as

$$\text{SI-RMSE} = \frac{1}{n}\sum_i R_i^2 - \frac{1}{n^2}\left(\sum_i R_i\right)^2. \tag{12}$$

For computing the multiscale gradient matching error, we adopt the same formula as in (11)

$$\text{MSGE} = \frac{1}{M}\sum_{k=1}^{K}\sum_{i=1}^{M}\left(|\nabla_x R_i^k| + |\nabla_y R_i^k|\right). \tag{13}$$

Compared with MAE and RMSE, the metrics SI-RMSE and MSGE are more concerned with the correctness of relative relationships in height maps, which are useful complementary metrics for the evaluation of transferable MHE models.

D. Experiments on GTAH Datasets

To verify the applicability and effectiveness of the proposed GTAH dataset and evaluate the proposed SDC module fairly, eight experiments are conducted on GTAH for performance comparison. First, following the work in [41], the U-Net [60] architecture with a CNN-based feature extraction backbone (ResNet-34) is adopted as a CNN-based baseline model for MHE. Then, Adabins [59] with ResNet-50 backbone and DenseViT [33] with the Vision Transformer backbone are selected as the state-of-the-art MDE methods for performance comparison. Furthermore, Swin-T is used as a Transformer-based feature extraction backbone for height estimation, which can be viewed as another baseline. To further explore the influence of the relative height loss functions on MHE, three different types of loss functions, $\mathcal{L}_{\text{msg}}$, $\mathcal{L}_{\text{si}}$, and $\mathcal{L}_r$, are added to constrain the relative relationship between pairwise pixels. The experimental results are provided in Table III.

When comparing the results of the two different baseline methods: U-Net and Swin-T, it is clear that the Transformer-based model significantly outperforms U-Net. Such results indicate that under the full-supervision setting within an unbiased dataset, the Transformer architecture can be more effective on the MHE task, benefiting from its excellent context modeling capability. Based on the Swin-T, the performance of $\mathcal{L}_{\text{msg}}$ is superior to the other two losses $\mathcal{L}_{\text{si}}$ and $\mathcal{L}_r$, which reveals that a reasonable relative height constraint is useful for improving the performance of MHE.

When integrating the proposed SDC module into the Swin-T, a notable performance gain is obtained, which verifies the effectiveness of the proposed SDC module for MHE due to its adaptive context modeling ability. It is worth noting that

TABLE VI

EXPERIMENTAL RESULTS ON THE JAX DATASET IN THE FEW-SHOT CROSS-DATASET TRANSFER SETTING. THE RESULTS OF USING 1% AND 5% TRAINING DATA ARE REPORTED. THE BEST RESULTS ARE IN BLUE, AND THE SECOND-BEST ONES ARE IN GREEN

Methods	Height Estimation Metrics (1% Training)				Height Estimation Metrics (5% Training)			
	MAE	SI-RMSE	RMSE	MSGE	MAE	SI-RMSE	RMSE	MSGE
U-Net (ImageNet) [55]	7.403	54.166	9.891	4.038	5.738	40.100	6.807	3.204
U-Net (GTAH) [55]	6.297	49.096	9.345	3.703	5.691	35.678	6.703	3.181
Adabins (ImageNet) [60]	6.510	48.972	9.001	3.970	5.22	35.239	6.592	3.202
Adabins (GTAH) [60]	5.779	40.335	7.327	3.522	5.082	27.418	6.390	3.092
DenseViT (ImageNet) [33]	7.165	50.660	9.777	3.975	5.676	35.543	6.534	3.144
DenseViT (GTAH) [33]	5.936	40.695	8.667	3.708	5.648	29.517	6.768	3.103
SwinUper (ImageNet) [56]	6.281	43.410	6.984	3.328	5.694	31.480	6.950	3.060
SwinUper (GTAH) [56]	5.597	40.314	6.600	3.324	4.739	25.276	6.602	2.920
SwinUper+$\mathcal{L}_{msg}$ (GTAH) [32]	5.645	40.719	6.818	3.420	4.828	26.054	6.632	2.998
SwinUper+SDC (GTAH) (Ours)	5.425	37.773	6.530	3.136	4.576	23.506	6.211	2.799

TABLE VII

EXPERIMENTAL RESULTS ON THE OMA DATASET IN THE FEW-SHOT CROSS-DATASET TRANSFER SETTING. THE RESULTS OF USING 1% AND 5% TRAINING DATA ARE REPORTED. THE BEST RESULTS ARE IN BLUE, AND THE SECOND-BEST ONES ARE IN GREEN

Methods	Height Estimation Metrics (1% Training)				Height Estimation Metrics (5% Training)			
	MAE	SI-RMSE	RMSE	MSGE	MAE	SI-RMSE	RMSE	MSGE
U-Net (ImageNet) [55]	3.755	17.682	6.399	2.031	3.475	19.402	6.156	1.925
U-Net (GTAH) [55]	3.572	16.452	6.388	2.009	3.439	14.815	6.028	1.873
Adabins (ImageNet) [60]	3.431	17.630	6.125	1.915	3.248	18.977	6.101	1.876
Adabins (GTAH) [60]	3.334	14.725	5.291	1.887	2.913	13.815	4.838	1.618
DenseViT (ImageNet) [33]	3.665	17.679	6.203	1.974	2.593	18.379	5.913	1.872
DenseViT (GTAH) [33]	3.546	15.257	5.619	1.898	3.250	11.009	5.490	1.650
SwinUper (ImageNet) [56]	3.780	16.350	4.549	1.980	3.577	14.757	4.465	1.786
SwinUper (GTAH) [56]	3.318	14.247	4.600	1.969	2.800	9.956	4.439	1.582
SwinUper+$\mathcal{L}_{msg}$ (GTAH) [32]	3.411	14.826	4.572	2.109	2.896	10.625	4.401	1.604
SwinUper+SDC (GTAH) (Ours)	3.237	13.714	4.480	1.866	2.735	9.519	4.238	1.536

there is great potential to combine $\mathcal{L}_{\text{msg}}$ with our proposed SDC, which may further boost MHE performance.

To further explore the generalization of the proposed method, we directly use the proposed method "SwinUper+SDC" that is pretrained on GTAH to predict the height of real-world images. Note that these images are randomly selected from Google Earth Map, as illustrated in Fig. 6. We can see that the pretrained model can reasonably estimate the height maps of these images, especially for the buildings with clear geometric information.

E. Experiments on Few-Shot Cross-Dataset Transfer

Few-shot learning has been studied heavily for image classification and semantic segmentation. However, for the dense regression task like MHE, there is still a lack of research on few-shot cross-dataset transfer. In this work, we fill in this gap by conducting cross-dataset transfer experiments from GTAH to the five real-world datasets under the few-shot setting. Before the few-shot cross-dataset transfer setting, we first conduct zero-shot transfer experiments to show the MHE results on real-city datasets using the GTAH pretrained weights with no fine-tuning process. The results in Table IV reveal that the MHE performance in the zero-shot transfer setting is poor due to significant domain shifts.

In the few-shot setting, for each dataset, only 1% or 5% of the training samples are randomly sampled for fast fine-tuning. As presented in Table II, for 1% setting, only less than 100 images are used for the model fine-tuning. The experimental results of AHN, JAX, OMA, ATL, and ARG are shown in Tables V–IX, respectively.

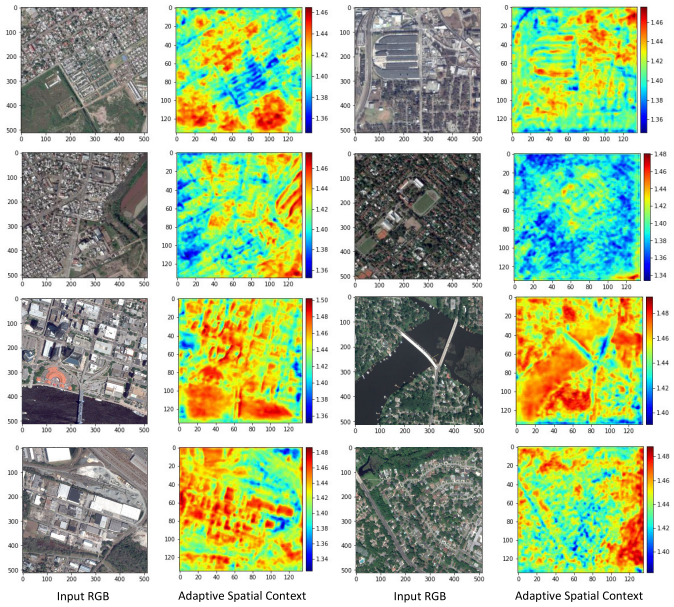


Fig. 7. Some visualization examples of the proposed SDC module. The SDC module can be observed to have learned to adjust the context information of different pixels in the cross-dataset transfer settings.

When we focus on the initialization strategy, it can be seen that the results of all the different models including U-Net, Adabins, DenseViT, and the SwinUper pretrained

TABLE VIII

EXPERIMENTAL RESULTS ON THE ATL DATASET IN THE FEW-SHOT CROSS-DATASET TRANSFER SETTING. THE RESULTS OF USING 1% AND 5% TRAINING DATA ARE REPORTED. THE BEST RESULTS ARE IN BLUE, AND THE SECOND-BEST ONES ARE IN GREEN

Methods	Height Estimation Metrics (1% Training)				Height Estimation Metrics (5% Training)			
	MAE	SI-RMSE	RMSE	MSGE	MAE	SI-RMSE	RMSE	MSGE
U-Net (ImageNet) [55]	13.693	143.72	12.985	8.446	12.503	129.958	17.871	7.384
U-Net (GTAH) [55]	13.354	146.156	12.884	7.893	10.160	123.002	18.863	6.228
Adabins (ImageNet) [60]	11.711	122.33	12.082	8.147	10.589	119.908	17.221	7.174
Adabins (GTAH) [60]	9.520	106.644	11.921	7.014	9.072	88.268	12.405	5.579
DenseViT (ImageNet) [33]	12.791	135.99	12.618	8.161	12.480	129.946	17.809	7.358
DenseViT (GTAH) [33]	10.296	119.817	11.612	7.167	9.385	129.759	13.784	7.105
SwinUper (ImageNet) [56]	13.827	143.518	12.732	8.199	10.370	99.236	12.566	6.057
SwinUper (GTAH) [56]	9.431	98.307	11.718	6.351	7.964	76.787	12.987	4.850
SwinUper+$\mathcal{L}_{msg}$ (GTAH) [32]	9.519	99.490	11.757	6.392	7.980	77.112	12.938	4.844
SwinUper+SDC (GTAH) (Ours)	9.386	98.738	11.605	6.307	7.959	77.427	11.913	4.836

TABLE IX

EXPERIMENTAL RESULTS ON THE ARG DATASET IN THE FEW-SHOT CROSS-DATASET TRANSFER SETTING. THE RESULTS OF USING 1% AND 5% TRAINING DATA ARE REPORTED. THE BEST RESULTS ARE IN BLUE, AND THE SECOND-BEST ONES ARE IN GREEN

Methods	Height Estimation Metrics (1% Training)				Height Estimation Metrics (5% Training)			
	MAE	SI-RMSE	RMSE	MSGE	MAE	SI-RMSE	RMSE	MSGE
U-Net (ImageNet) [55]	3.658	12.561	8.484	2.497	3.392	12.168	8.351	2.398
U-Net (GTAH) [55]	3.611	13.604	8.546	2.581	3.091	10.217	8.121	2.146
Adabins (ImageNet) [60]	3.621	12.551	7.007	2.486	3.208	11.089	6.693	2.099
Adabins (GTAH) [60]	3.465	11.365	6.719	2.388	2.904	9.712	6.211	1.942
DenseViT (ImageNet) [33]	3.632	12.389	8.261	2.485	3.173	11.998	7.703	2.311
DenseViT (GTAH) [33]	3.520	11.393	7.452	2.395	2.972	10.053	4.997	1.941
SwinUper (ImageNet) [56]	3.742	12.516	6.340	2.523	2.894	9.548	6.107	2.004
SwinUper (GTAH) [56]	3.486	11.976	6.260	2.508	2.831	9.015	6.048	1.910
SwinUper+$\mathcal{L}_{msg}$ (GTAH) [32]	3.497	11.910	6.193	2.474	2.926	9.774	6.096	2.038
SwinUper+SDC (GTAH) (Ours)	3.458	10.467	6.187	2.230	2.786	8.883	6.030	1.804

on GTAH have a dramatic superiority to those pretrained on ImageNet. Especially for the ATL dataset with higher height distribution, ImageNet pretrained U-Net and Swin-T experience a performance collapse, whereas these models pretrained on GTAH maintain a stable performance. The results demonstrate that our proposed GTAH dataset is more suitable for MHE initialization.

Next, the results of the Swin-T model pretrained on GTAH with relative constraint loss $\mathcal{L}_{\text{msg}}$ show that the introduction of $\mathcal{L}_{\text{msg}}$ is not beneficial for improving performance in general. The reason may be that introducing the relative constraint by loss functions is not useful for improving the generalizability of the MHE model across different datasets. In contrast, the proposed SDC module is still effective in the few-shot setting. The model "SwinUper + SDC (GTAH)" obtains the best results on all the datasets. Compared with the baseline method "SwinUper (GTAH)," the proposed model "SwinUper + SDC (GTAH)" shows virtue of its overall superiority with a considerable performance gain, which verifies that the adaptive scale modeling ability is helpful for the Swin-T by learning an adaptive receptive field.

To intuitively illustrate the effect of the SDC module, some visualization examples of the dynamic spatial context are provided in Fig. 7. In Fig. 7, we can see that the low-frequency region needs a larger receptive field to acquire enough context information, while the high-frequency region only requires a relatively smaller receptive field for the MHE task.

Some visualization examples on the five real-world datasets under the few-shot transfer setting are presented in Fig. 8.

F. ImageNet Pretraining Versus GTAH Pretraining

To further study the effectiveness of model pretraining in the cross-dataset transfer setting, we present and analyze the loss trends of the model "SwinUper+SDC" during the few-shot model fine-tuning stage. As shown in Fig. 9, from the loss trend we can see that models with GTAH pretrained parameters can converge faster. Especially on the JAX, OMA, and ATL datasets, models initialized with the ImageNet pretrained parameters are difficult to converge.

1) Visualization of the Loss Tendency: In Fig. 9, we can observe that for the datasets of AHN and ARG, both ImageNet and GTAH pretrained parameters can accelerate model training, while using GTAH can result in a faster convergence rate at the early stage. Taking into consideration their test results in Tables V and IX, the Swin-T pretrained on the GTAH dataset still outperforms that pretrained on ImageNet. When we turn to the JAX and ATL datasets, the model using GTAH pretrained parameters experiences a rapid decline and shows a dramatic advantage to ImageNet. From the perspective of loss tendency, our proposed GTAH dataset can facilitate the training process of all the datasets, albeit to different extents.

2) Visualization of the Weight Distribution: We also visualize and compare the weight distributions of deep models trained on ImageNet, GTAH, and real-world datasets, respectively. As presented in Fig. 10, the weight distribution of the last two layers (Layer1 and Layer2) of the SwinUper method is visualized. For both the layers, the parameters

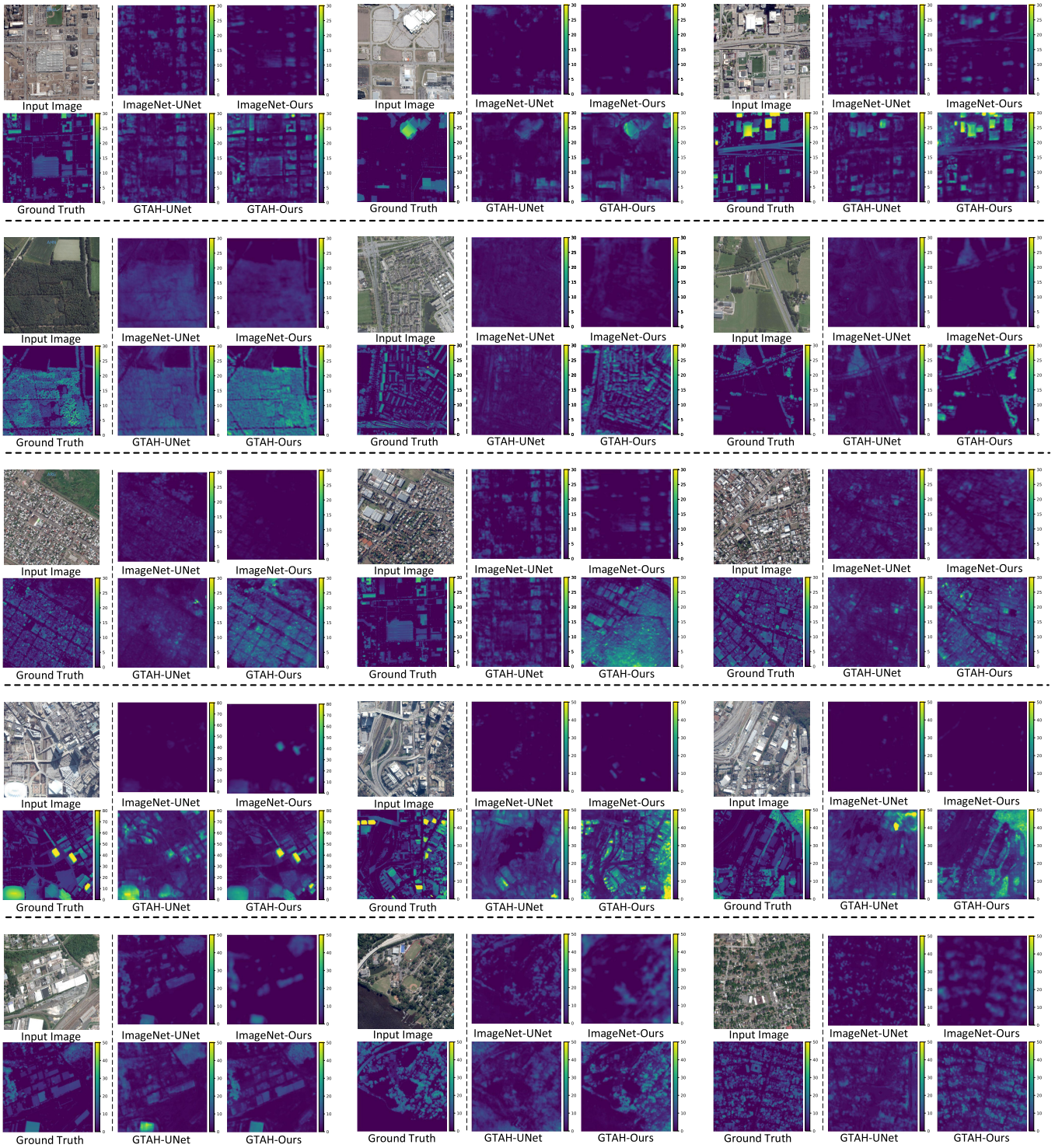


Fig. 8. Visualization results of height estimation in a few-shot **(1% of the training data)** cross-dataset transfer setting. In this figure, the predicted height maps of five real-world datasets OMA, AHN, ARG, ATL, and JAX are shown. It can be seen that our method "SwinUper+SDC (GTAH)" can obtain much better results than other methods in the few-shot across-dataset transfer setting.

trained on GTAH have a more similar distribution to those trained on real-world datasets than the ImageNet pretrained parameters. Especially for Layer1, the distribution of weights trained on GTAH is highly consistent with those trained on real datasets. For Layer2, the differences in weight distributions between ImageNet and real-world datasets become larger. This is reasonable because the shallow layers mainly extract the universal representations, while the final layer is usually responsible for the dataset-specific predictions.

3) Comparison With Pretrained Weights on Remote Sensing Data: In Table X, we also compare the performance of weights pretrained on GTAH with other remote sensing datasets. Considering that natural images in ImageNet are different from remotely sensed images, we compare our

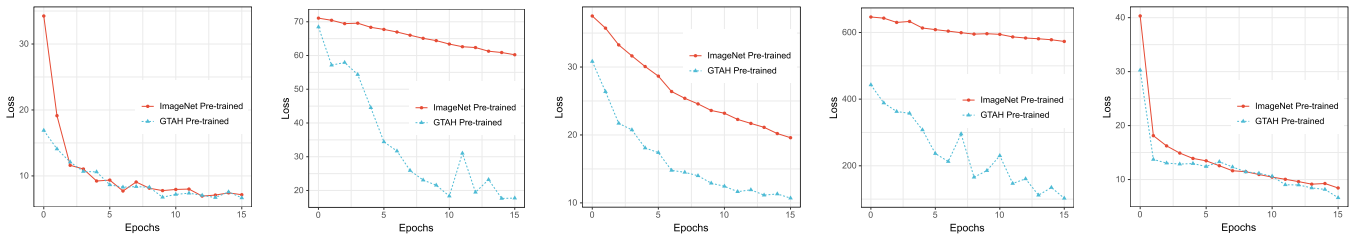


Fig. 9. Training loss of the "Swin-SDC" (Swin-T followed by the SDC module) on five real-world MHE datasets after being pretrained on ImageNet and GTAH. (a) AHN. (b) JAX. (c) OMA. (d) ATL. (e) ARG.

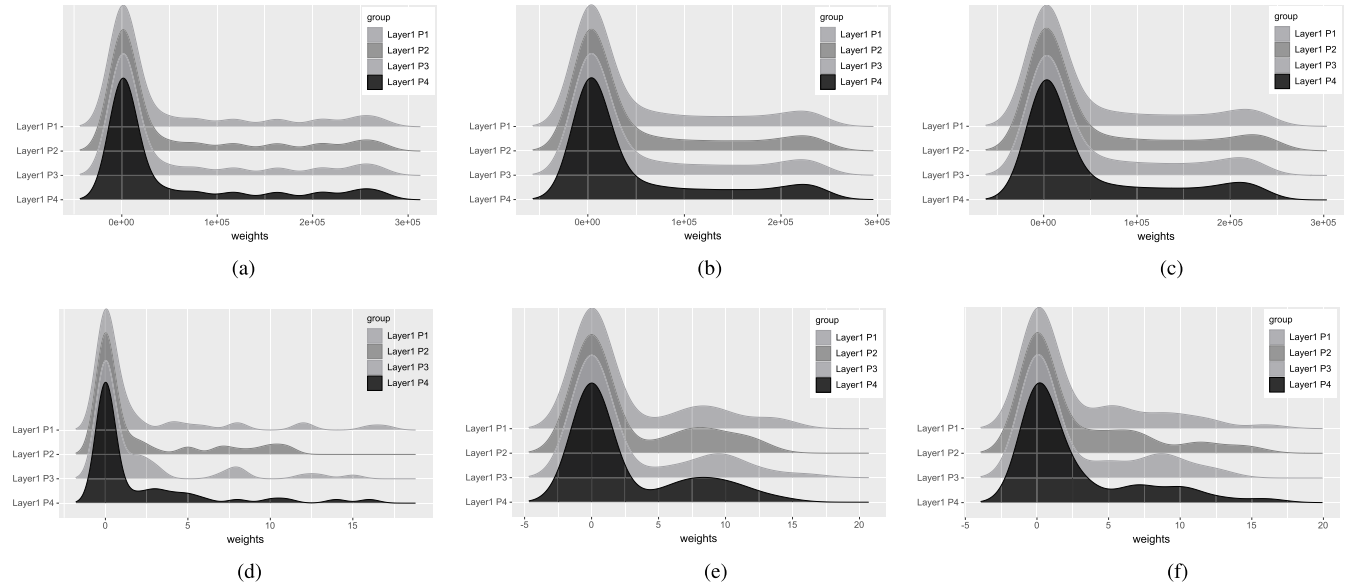


Fig. 10. Weight distributions of the Swin-T trained on ImageNet, GTAH, and the combination of real-world datasets, respectively. (a) Layer1-ImageNet. (b) Layer1-GTAH. (c) Layer1-Real. (d) Layer2-ImageNet. (e) Layer2-GTAH. (f) Layer2-Real.

TABLE X

EXPERIMENTAL RESULTS ON THE AHN DATASET WITH DIFFERENT PRETRAINING METHODS. THE BEST RESULTS ARE IN BLUE, AND THE SECOND-BEST ONES ARE IN GREEN

Methods	Height Estimation Metrics (1% Training)				Height Estimation Metrics (5% Training)			
	MAE	SI-RMSE	RMSE	MSGE	MAE	SI-RMSE	RMSE	MSGE
SwinUper (ImageNet) [56]	2.394	9.002	5.694	1.724	2.104	9.297	5.585	1.435
SwinUper (SSLTransformerRS) [62]	2.347	8.626	5.715	1.613	2.161	9.055	5.674	1.423
SwinUper (GTAH) [56]	2.156	8.303	5.564	1.455	2.107	9.021	5.569	1.426
SwinUper+SDC (GTAH) (Ours)	2.117	8.285	5.530	1.441	2.001	8.981	5.439	1.417

method with SSLTransformerRS [61] that are pretrained on the Sentinel-2 dataset using self-supervised learning. As shown in Table X, SwinUper (SSLTransformerRS) is slightly better than SwinUper (ImageNet). Pretraining using GTAH can achieve superior performance than others, which indicates the effectiveness of the proposed dataset.

VI. CONCLUSION

In this article, we study the transferability of height estimation models in a cross-dataset transfer setting. To start with, a new large-scale synthetic dataset, named GTAH, for height estimation from monocular remote sensing images is constructed and released. GTAH contains highly accurate high-resolution RGB/height image pairs captured under different imaging conditions, which can be helpful to foster research on MHE. Furthermore, we also collect and release a large-scale real-world dataset termed AHN, for the MHE task. Then, we study the transferability of deep learning models for MHE in a cross-dataset setting, which is more consistent with real-world applications. To achieve this goal, a large-scale benchmark dataset for cross-dataset transfer learning on the MHE task is constructed. Furthermore, a new experimental protocol, few-shot cross-dataset transfer, is designed to evaluate the generalizability of MHE models in a cross-dataset setting. In addition, an SDC module is designed to handle the severe scale variation problem. The experimental results have verified the effectiveness of the proposed new datasets and methods for the height estimation task.

REFERENCES

[1] L. Mou and X. X. Zhu, "IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," 2018, *arXiv:1802.10249*.

[2] R. R. Navalgund, V. Jayaraman, and P. S. Roy, "Remote sensing applications: An overview," *Current Sci.*, vol. 93, no. 12, pp. 1747–1766, Dec. 2007.

[3] I. Masser, "Managing our urban future: The role of remote sensing and geographic information systems," *Habitat Int.*, vol. 25, no. 4, pp. 503–512, Dec. 2001.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[5] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" 2014, *arXiv:1411.1792*.

[6] Rockstar. (2021). *Grand Theft Auto V*. [Online]. Available: https://www.rockstargames.com

[7] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 18, 2005, pp. 1–8.

[8] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: Learning 3D scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2008.

[9] J. Konrad, M. Wang, and P. Ishwar, "2D-to-3D image conversion by learning depth from examples," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 16–22.

[10] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2650–2658.

[11] S. Zhu, G. Brazil, and X. Liu, "The edge of depth: Explicit constraints between segmentation and depth," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13113–13122.

[12] J. Jiao, Y. Cao, Y. Song, and R. Lau, "Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 53–69.

[13] S. M. H. Miangoleh, S. Dille, L. Mai, S. Paris, and Y. Aksoy, "Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9680–9689.

[14] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 740–756.

[15] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 270–279.

[16] A. Pilzer, S. Lathuilière, N. Sebe, and E. Ricci, "Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9760–9769.

[17] R. Peng, R. Wang, Y. Lai, L. Tang, and Y. Cai, "Excavating the potential capacity of self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15540–15549.

[18] H. Jung, E. Park, and S. Yoo, "Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12622–12632.

[19] Y. Jiao, T. D. Tran, and G. Shi, "EffiScene: Efficient per-pixel rigidity inference for unsupervised joint learning of optical flow, depth, camera pose and motion segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5534–5543.

[20] F. Tosi et al., "Distilled semantics for comprehensive scene understanding from videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4653–4664.

[21] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3828–3838.

[22] C. Shu, K. Yu, Z. Duan, and K. Yang, "Feature-metric loss for self-supervised learning of depth and egomotion," in *Proc. Eur. Conf. Comput. Vis.*, Nov. 2020, pp. 572–588.

[23] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3D packing for self-supervised monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2485–2494.

[24] X. Lyu et al., "HR-Depth: High resolution self-supervised monocular depth estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 3, pp. 2294–2301.

[25] V. Guizilini, R. Hou, J. Li, R. Ambrus, and A. Gaidon, "Semantically-guided representation learning for self-supervised monocular depth," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–14.

[26] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2015.

[27] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 239–248.

[28] P.-Y. Chen, A. H. Liu, Y.-C. Liu, and Y. F. Wang, "Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2619–2627.

[29] L. Wang, J. Zhang, O. Wang, Z. Lin, and H. Lu, "SDC-Depth: Semantic divide-and-conquer network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 538–547.

[30] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct./Nov. 2019, pp. 5684–5693.

[31] J.-H. Lee and C.-S. Kim, "Multi-loss rebalancing algorithm for monocular depth estimation," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, Aug. 2020, pp. 785–801.

[32] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, Mar. 2022, doi: 10.1109/TPAMI.2020.3019967.

[33] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12179–12188.

[34] Z. Li et al., "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6177–6186.

[35] S. Srivastava, M. Volpi, and D. Tuia, "Joint height estimation and semantic labeling of monocular aerial images with CNNS," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 5173–5176.

[36] P. Ghamisi and N. Yokoya, "IMG2DSM: Height simulation from single imagery using conditional generative adversarial net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 794–798, May 2018.

[37] M. E. Paoletti, J. M. Haut, P. Ghamisi, N. Yokoya, J. Plaza, and A. Plaza, "U-IMG2DSM: Unpaired simulation of digital surface models with generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 7, pp. 1288–1292, Jul. 2021.

[38] H. A. Amirkolaee and H. Arefi, "CNN-based estimation of pre- and post-earthquake height models from single optical images for identification of collapsed buildings," *Remote Sens. Lett.*, vol. 10, no. 7, pp. 679–688, Jul. 2019.

[39] H. A. Amirkolaee and H. Arefi, "Height estimation from single aerial images using a deep convolutional encoder–decoder network," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, pp. 50–66, Mar. 2019.

[40] W. Liu, W. Zhang, X. Sun, Z. Guo, and K. Fu, "HECR-Net: Height-embedding context reassembly network for semantic segmentation in aerial images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9117–9131, 2021.

[41] G. Christie, R. R. R. M. Abujder, K. Foster, S. Hagstrom, G. D. Hager, and M. Z. Brown, "Learning geocentric object pose in oblique monocular images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14512–14520.

[42] J. Mahmud, T. Price, A. Bapat, and J.-M. Frahm, "Boundary-aware 3D building reconstruction from a single overhead image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 441–451.

[43] L. Madhuanand, F. Nex, and M. Y. Yang, "Self-supervised monocular depth estimation from oblique UAV videos," *ISPRS J. Photogramm. Remote Sens.*, vol. 176, pp. 1–14, Jun. 2021.

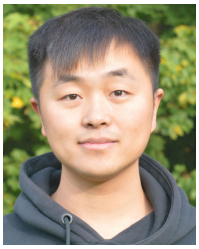
[44] Z. Zheng, Y. Zhong, and J. Wang, "Pop-Net: Encoder-dual decoder for semantic segmentation and single-view height estimation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 4963–4966.

[45] C.-J. Liu, V. A. Krylov, P. Kane, G. Kavanagh, and R. Dahyot, "IM2ELEVATION: Building height estimation from single-view aerial imagery," *Remote Sens.*, vol. 12, no. 17, p. 2719, Aug. 2020.

[46] X. Li, M. Wang, and Y. Fang, "Height estimation from single aerial images using a deep ordinal regression network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[47] T. Van Dijk and G. De Croon, "How do neural networks see depth in single images?" in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2183–2191.

[48] G. Christie, K. Foster, S. Hagstrom, G. D. Hager, and M. Z. Brown, "Single view geocentric pose in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1162–1171.
[49] M. Bosch, K. Foster, G. Christie, S. Wang, G. D. Hager, and M. Brown, "Semantic stereo for incidental satellite images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1524–1532.
[50] S. Kunwar et al., "Large-scale semantic 3-D reconstruction: Outcome of the 2019 IEEE GRSS data fusion contest—Part A," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 922–935, 2021.
[51] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.
[52] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2366–2374.
[53] W. Chen, Z. Fu, D. Yang, and J. Deng, "Single-image depth perception in the wild," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 730–738.
[54] C. Wang, S. Lucey, F. Perazzi, and O. Wang, "Web stereo video supervision for depth prediction from dynamic scenes," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2019, pp. 348–357.
[55] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.
[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[57] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 418–434.
[58] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in Adam," 2018.
[59] S. F. Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth estimation using adaptive bins," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jan. 2021, pp. 4009–4018.
[60] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[61] L. Scheibenreif, J. Hanna, M. Mommert, and D. Borth, "Self-supervised vision Transformers for land-cover segmentation and classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1422–1431.



Zhitong Xiong (Member, IEEE) received the Ph.D. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 2021.

He is currently a Research Scientist and leads the ML4Earth Working Group, Chair of Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany. His research interests include computer vision, machine learning, Earth observation, and Earth system modeling.

Wei Huang received the B.E. and M.E. degrees from Northwestern Polytechnical University, Xi'an, China, in 2018 and 2021, respectively.

He is currently a Visiting Student with the Chair of Data Science in Earth Observation, Technical University of Munich, Munich, Germany. His research interests include computer vision and remote sensing.

Jingtao Hu received the M.E. degree in control engineering from the Lanzhou University of Technology, Lanzhou, China, in 2018. He is currently pursuing the Ph.D. degree with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China.

His research interests include remote sensing, computer vision, and machine learning.

Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.-Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

Since 2019, she has been a Co-Coordinator of the Munich Data Science Research School (www.mu-ds.de), TUM. Since 2019, she has also been heading the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Munich, Space and Transport." Since May 2020, she has been the Director of the International Future AI Laboratory "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. Since October 2020, she has also been serving as a Co-Director of the Munich Data Science Institute (MDSI), TUM. She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy; Fudan University, Shanghai, China; The University of Tokyo, Tokyo, Japan; and the University of California at Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. She is currently a Professor with the Chair of Data Science in Earth Observation, Technical University of Munich (TUM), and the Head of the Department "EO Data Science" at the Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. She is also a Visiting AI Professor at the ESA's Phi-lab, Frascati, Itlay. Her main research interests include remote sensing and Earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.