Self-Supervised Pretraining With Monocular Height Estimation for Semantic Segmentation

Zhitong Xiong, *Member, IEEE*, Sining Chen, Yilei Shi, *Member, IEEE*, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*— Monocular height estimation (MHE) is key for generating 3-D city models, essential for swift disaster response. Moving beyond the traditional focus on performance enhancement, our study breaks new ground by probing the interpretability of MHE networks. We have pioneeringly discovered that neurons within MHE models demonstrate selectivity for both height and semantic classes. This insight sheds light on the complex inner workings of MHE models and inspires innovative strategies for leveraging elevation data more effectively. Informed by this insight, we propose a pioneering framework that employs MHE as a self-supervised pretraining method for remote sensing (RS) imagery. This approach significantly enhances the performance of semantic segmentation tasks. Furthermore, we develop a disentangled latent transformer (DLT) module that leverages explainable deep representations from pretrained MHE networks for unsupervised semantic segmentation. Our method demonstrates the significant potential of MHE tasks in developing foundation models for sophisticated pixel-level semantic analyses. Additionally, we present a new dataset designed to benchmark the performance of both semantic segmentation and height estimation tasks. The dataset and code will be publicly available at https://github.com/zhu-xlab/DLT-MHE.pytorch.**

***Index Terms*— Foundation models, interpretable deep learning, monocular height estimation (MHE), self-supervised pretraining.**

I. Introduction

THE geometric information of 3-D cities can be useful for urban planning, damage monitoring, disaster forecasting, and so on. In this context, obtaining geometric information efficiently from remote sensing (RS) imagery is essential for a rapid response to time-critical world events, e.g., natural hazards and damage assessment [1], [2].

Motivated by the development of monocular depth estimation (MDE) task [5], [6], [7], various methods have been proposed for monocular height estimation (MHE). Mou and Zhu [8] designed a residual convolutional network for height estimation and demonstrated its effectiveness on the instance segmentation task. Christie et al. [9], [10] proposed to estimate the geocentric pose from monocular oblique images. Complementary to optical data, Sun et al. [11] proposed a workflow for building height retrieval in single SAR images. Since estimating height from monocular images is an ill-posed task, improving the interpretability and reliability of deep models is crucial for risk-sensitive applications. However, few works focus on understanding deep networks for MHE.

As for interpretable deep models [12], previous research on image classification [13], [14] and object detection [15] has been studied. However, MHE is a dense prediction task involving pixel-wise regression. This makes image-level and object-level interpretation methods not applicable. The most relevant task to MHE is MDE. Dijk and Croon [16] investigated important visual cues in input monocular images for depth estimation. Hu et al. [17] tried to find the most relevant sparse pixels for estimating the depth. An interesting work from [18] first found the depth selectivity of some hidden neurons, which showed promising insights for interpreting MDE models.

Although MHE shares some similar characteristics with MDE, several aspects make them quite different. First, height is an inherent attribute of objects, which should not change under different views. However, the depth of objects highly depends on the camera pose. Second, estimating height from the top view may be far more ambiguous than depth estimation from the street view because of the severe occlusion and lack of textures. Third, object types, scales, and scene layouts vary greatly in remotely sensed imagery. These differences also make it unsuitable to apply interpretation methods of MDE to MHE.

Considering this issue, in this study, we propose to explain the deep MHE model by a network dissection to explore the properties of hidden neurons and understand what internal representations the MHE model has learned. From observation, we find that MHE networks learn disentangled representations to different semantic concepts and height ranges. As shown in Fig. 1, roads, trees, and buildings are automatically recognized by some neurons. Some of the neurons are selective to different height ranges. Based on the finding of semantic selectivity, we introduce a framework that utilizes MHE as a self-supervised pretraining method tailored for RS imagery.

Manuscript received 27 March 2024; accepted 30 April 2024. Date of current version 22 July 2024. This work was supported 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; 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 Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV) Based on a Resolution of the German Bundestag under Grant 67KI32002B, Acronym: EKAPEx; in part by European Commission through the Project "ThinkingEarth—Copernicus Foundation Models for a Thinking Earth" through the Horizon 2020 Research and Innovation Program under Agreement 101130544; and in part by Munich Center for Machine Learning. *(Corresponding author: Xiao Xiang Zhu.)*

Zhitong Xiong and Sining Chen are with the Chair of Data Science in Earth Observation, Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: zhitong.xiong@tum.de; sining.chen@tum.de).

Yilei Shi is with the School of Engineering and Design, Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: yilei.shi@tum.de).

Xiao Xiang Zhu is with the Chair of Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany, and also with Munich Center for Machine Learning, 80333 Munich, Germany (e-mail: xiaoxiang.zhu@tum.de).

Digital Object Identifier 10.1109/TGRS.2024.3412629

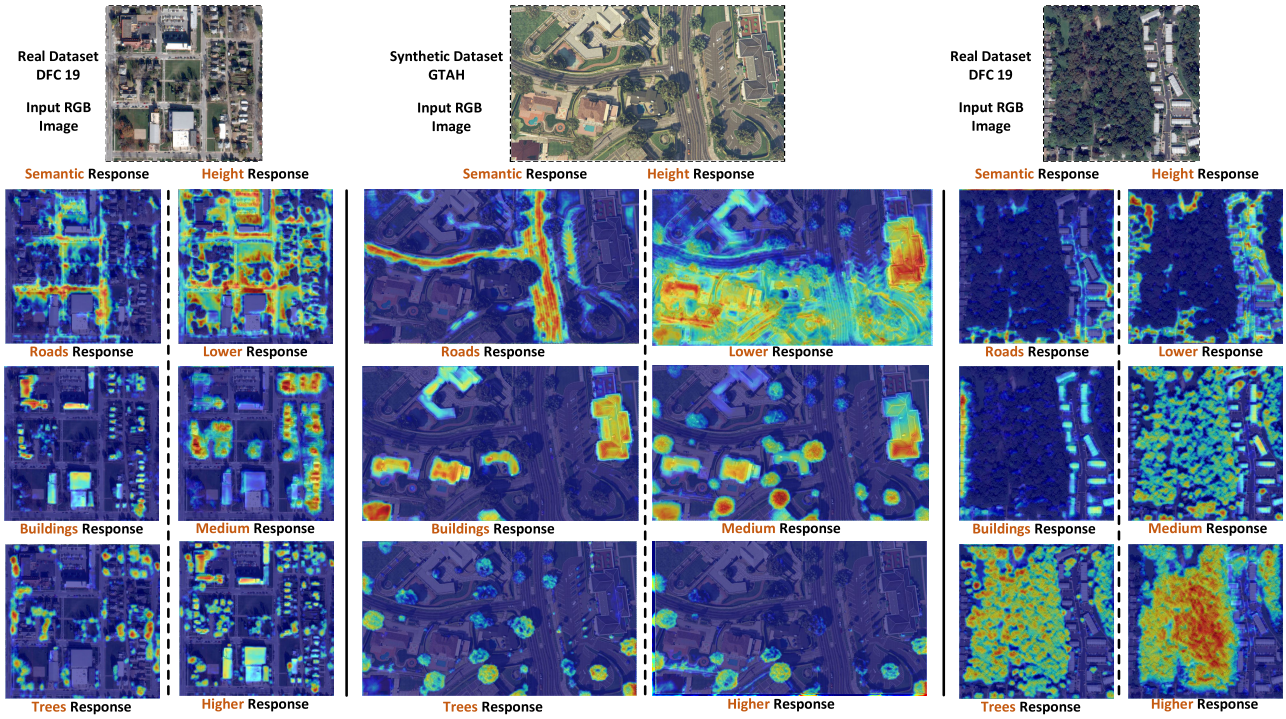


Fig. 1. MHE networks learn to recognize different semantic objects (road, building, and tree) and height ranges implicitly. This figure shows the strong selectivity of Transformer-based MHE networks on both the GTAH dataset [3] and the DFC 2019 [4] dataset.

By adopting the pretrained weights to downstream semantic segmentation tasks, the performance can be substantially enhanced. This highlights the potential of incorporating MHE in designing RS foundation models [19] for pixel-level tasks.

While multimodal pretraining has been extensively explored in existing [20], with approaches like CROMA [21] leveraging SAR and optical image pairs for cross-modal reconstruction and contrastive learning, and DeCUR [22] focusing on distinguishing between modality-common and modality-specific representations, our work focuses more on the elevation data and introduces significant innovations and advantages. First, we utilize explainable deep learning to elucidate the underlying reasons why MHE is an effective self-supervised pretraining method for enhancing semantic analysis tasks. Second, our utilization of the MHE task for pretraining endows the resulting model with dual functionalities: not only does it excel in downstream semantic analysis tasks, but it also operates effectively as a height estimation network. This bifunctionality of the pretrained model highlights our contribution: leveraging MHE to bridge the gap between semantic analysis and height estimation in RS.

Furthermore, we demonstrate that the pretrained networks for MHE can be adapted to unsupervised semantic segmentation. To this end, we introduce a disentangled latent transformer (DLT) module that adeptly refines and constrains interim representations to achieve semantic coherence. Consequently, the model facilitates simultaneous height estimation and semantic segmentation tasks, leveraging solely height data for supervision. This methodology not only delineates a novel pathway for integrating height estimation with semantic segmentation but also surpasses the performance set by existing unsupervised segmentation methods that utilize both the RGB and height data.

To contribute more datasets containing both height maps and semantic labels, we construct the Washington DC (WDC) dataset, to foster research on improving both the interpretability and performance of MHE models. Our contributions can be summarized as follows.

1) *Neuron Selectivity Insights:* We identify that neurons within pretrained MHE models demonstrate pronounced selectivity toward both height ranges and semantic classes. This discovery is instrumental in guiding future research to optimize the utilization of elevation data in RS applications.
2) *Self-Supervised Pretraining:* Based on the insights into semantic selectivity, we introduce a self-supervised pretraining framework based on MHE. This approach significantly enhances the performance of semantic segmentation tasks, showcasing the utility of pretrained weights derived from MHE models.
3) *DLT:* We propose a novel DLT module to learn explainable and effective deep representations for joint segmentation and height estimation. This module can distill explainable deep representations from pretrained MHE networks, facilitating unsupervised semantic segmentation tasks.
4) *New RGB-H Dataset:* A new dataset containing RGB and normalized digital surface models (nDSM) is constructed and released to evaluate joint semantic segmentation and height estimation.

II. Related Work

A. Interpretable and Explainable Deep Neural Networks

Interpretable deep models and explainable deep networks both aim to make the complex decision-making processes of deep learning transparent, albeit through different approaches.

On the one hand, interpretable models are inherently designed to be transparent, allowing for a direct understanding of their internal operations. On the other hand, explainable networks rely on external methods or frameworks to make sense of their more complex decisions after the fact.

Several methods attempted to design inherently interpretable models for image classification and object detection tasks. Chen et al. [14] proposed to find prototypical parts and explain the reason for making final decisions. Toward the interpretability for the person reidentification task, Liao and Shao [23] designed a model to make the matching process of feature maps explicit. Zhang et al. [13] designed interpretable CNNs by making each filter represent a specific object part. Liang et al. [24] trained interpretable CNNs by learning class-specific deep filters, encouraging each filter only to account for a few classes. Similarly, You et al. [18] proposed to improve depth selectivity by designing specific loss functions for MDE models. To understand what these MDE networks have learned, [16] studied important visual clues used by deep networks when predicting depth. They mainly focused on the object-level interpretation of MDE networks. [17] attempted to find only a selected sparse set of image pixels to estimate depth. A separate network is designed to predict those sparse pixels. However, these methods neglect the inherent representations that the model has learned.

Many researchers focused on saliency-based and attribution-based methods [12], [25], [26] for explainable deep networks. They aimed to highlight which pixels of input images are important for predicting the result [27], [28]. However, attribution and saliency-based methods are not directly applicable to dense prediction tasks, including MDE and MHE, since it is not reasonable to highlight all the pixels to attribute the dense prediction globally. Gu and Dong [29] proposed a local attribution method for interpreting super-resolution networks. They chose to interpret features instead of pixels for super-resolution tasks, which inspires our work on pixel-wise attribution for MHE networks.

B. Monocular Height Estimation

With the development of deep learning, various methods have been proposed for MHE. Srivastava et al. [30] proposed to predict height and semantic labels jointly in a multi-task deep learning framework. Mou and Zhu [8] designed a residual CNN for height estimation and demonstrated its effectiveness on instance segmentation tasks. Besides, conditional generative adversarial network (cGAN) Ghamisi and Yokoya [31] was proposed to frame height estimation as an image translation task. Kunwar et al. [32] exploited semantic labels as priors to enhance the performance of height estimation on the large-scale urban semantic 3-D (US3D) dataset [33]. Xiong et al. [3] designed and constructed a large-scale benchmark dataset for cross-dataset transfer learning on the height estimation task, which includes a large-scale synthetic dataset and several real-world datasets. Swin transformer [34] was used in this work for transferable representation learning on the MHE task. HTC-dc Net [35] introduces a classification–regression paradigm for MHE instead of formalizing the problem as a regression task. The experimental results show its superiority over existing methods.

C. Self-Supervised Pretraining on Multimodal RS Data

Self-supervised pretraining [36], [37], [38] on RS imagery can largely enhance the performance of deep learning models in various RS tasks. For multimodal RS data, CROMA [21] designs two unimodal encoders to encode both multispectral and SAR data individually. Then, a cross-modal radar-optical transformer is proposed to extract the unified representation. DeCUR [22] develops a self-supervised foundation model that decouples the unique and common representations between two different modalities. SpectralGPT [39] is a masked autoencoder-based foundation model designed for hyperspectral RS data to preserve spectral characteristics. Distinct from existing models, we innovatively adopt MHE as a self-supervised pretraining strategy. This approach is straightforward and effective, endowing the pretrained models with pixel-level semantic representations. Our findings illuminate the potential of elevation data in developing RS foundational models for pixel-level tasks.

III. Methodology

To understand the behavior of learned hidden representations, we examine the learned knowledge of MHE networks by visualizing the deep representations and computing the semantic and height selectivity scores. Based on the findings, two contributions are made in this work: 1) we propose an MHE-based pretraining framework for semantic segmentation and 2) we design a novel DLT module for joint segmentation and height estimation using only the supervision from height data. The whole architecture is illustrated in Fig. 2. In this section, we will introduce them in detail.

A. Network Dissection for MHE Models

To estimate the height of an object, given a top-view image, human beings tend to recognize the object first. This cognitive process of humans highly inspires us to explore the learned deep representations of MHE models. Objects of different semantic types usually have distinct height attributes. Thus, the geometric information in height maps should have a high correlation with the semantic information [32]. Motivated by this, we choose to examine the learned interior activations to find human-understandable patterns.

From numerous visualizations of deep activations, we find that some neurons of the deep networks are highly selective to different semantic classes. As shown in Fig. 1, different semantic objects including roads, buildings, and trees are localized accurately in an implicit manner with only height supervision. This finding supports the assumption that semantic information and geometric information are highly correlated. Moreover, we also find that some neurons are selective to different height ranges, which shares the same conclusion with [18] for the MDE task.

To quantify the selectivity of the network neurons, we compute both the class-selectivity [40] and the height-selectivity [18] for each internal unit. In this work, we utilize the Swin transformer [34] as the backbone for the height estimation task. Denote each test sample in the dataset D as $(\mathbf{x_i}, \mathbf{y_i}, \mathbf{h_i})$, where $\mathbf{x_i} \in \mathbb{R}^{3\times H\times W}$ is the input image, $\mathbf{y_i} \in \mathbb{R}^{H\times W}$ and $\mathbf{h_i} \in \mathbb{R}^{H\times W}$ are the ground truth semantic map and height map, respectively. $i \in \{1, \ldots, N\}$. Note that the ground truth

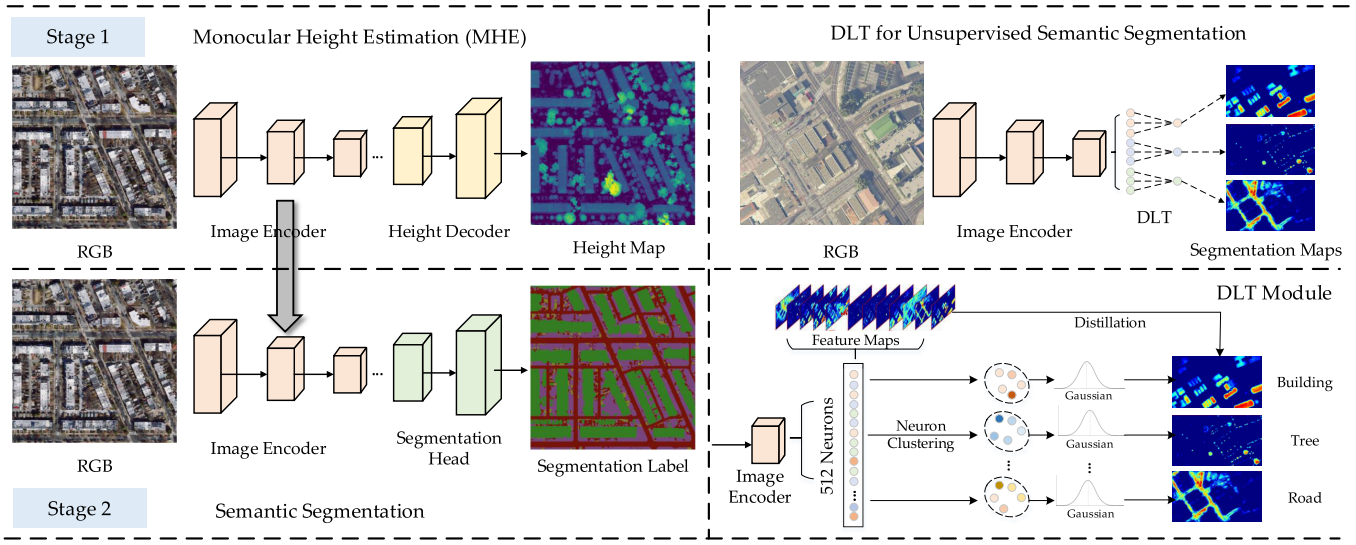


Fig. 2. Illustration of the proposed method. (Left) Shows the MHE pretraining for semantic segmentation. In stage one, we pretrain the model for height estimation. The pretrained weights of the image encoder are transferred to semantic segmentation tasks in stage two. (Right) Presents the proposed DLT module, which enables unsupervised semantic segmentation only using the pretrained image encoder in stage one.

class label $\mathbf{y_i}$ is only used for the selectivity score computing, and not used for the height estimation task. We upsample the k_{th} feature map of the penultimate layer to the image size as $F_k(x_i) \in \mathbb{R}^{H\times W}$. The class-selectivity and height-selectivity can be computed by

$$\mathrm{CR}_k^c = \frac{\sum_1^N S\left(F_k(\mathbf{x_i}) \odot M_i^c\right)}{\sum_1^N S\left(M_i^c\right)}; \quad \mathrm{HR}_k^h = \frac{\sum_1^N S\left(F_k(\mathbf{x_i}) \odot M_i^h\right)}{\sum_1^N S\left(M_i^h\right)}$$
$$\mathrm{CS}_k = \frac{|\mathrm{CR}_k^{\max} - \bar{\mathrm{CR}}_k^{-\max}|}{|\mathrm{CR}_k^{\max} + \bar{\mathrm{CR}}_k^{-\max}|}; \quad \mathrm{HS}_k = \frac{|\mathrm{HR}_k^{\max} - \bar{\mathrm{HR}}_k^{-\max}|}{|\mathrm{HR}_k^{\max} + \bar{\mathrm{HR}}_k^{-\max}|} \tag{1}$$

where CR_k^c and HR_k^h are the average response of the neurons for different classes and height ranges. h is the index of discretized height ranges and c is the index of semantic classes. M^c is a binary mask indicating the pixels with semantic class c. M^h is also a binary mask indicating the pixels in height range h. $S(\cdot)$ denotes the summation operation on all elements in a matrix. We use $\odot$ to represent the element-wise multiplication. To ensure the stability of the computation and avoid potential division by zero errors, we incorporate a small epsilon value ($\epsilon = 1e-6$) into the denominators when computing the average response of neurons in practice. With the defined average responses CR_k^c and HR_k^h, class-selectivity and height-selectivity for the k_{th} neuron CS_k and HS_k can be computed.

In Fig. 3, the height range selectivity (first row) and class selectivity (second row) of the Transformer-based MHE model are visualized. The selectivity scores for the height range of neurons 277, 88, and 59 are larger than 0.85. We can see that these neurons are highly selective to low, medium, and high height ranges. The class-selectivity of neurons 255, 87, and 20 are larger than 0.8, and they are highly selective to categories *Ground, High Vegetation,* and *Building*, respectively. However, there are no clear selective neurons for *Water* and *Elevated Road*, since these categories have no clear differences compared with *Ground* regarding height

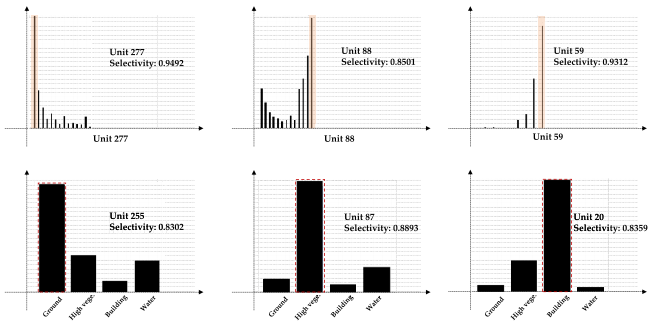


Fig. 3. Visualization of the average response for feature map activations with large height range selectivity (first row) and large class selectivity (second row) on the DFC 2019 dataset [4].

changes. The finding that some neurons are selective to different height ranges shares a similar conclusion with [18] for MDE tasks. To provide a clearer presentation of the high selectivity score, we visualize the activation maps and the discretized height maps in three different height ranges in Fig. 4. Specifically, we first find the max height h_m value for each height map. Then, we choose three height ranges by using different thresholds: 1) lower than $0.2 \cdot h_m$; 2) the range between $0.2 \cdot h_m$ and $0.7 \cdot h_m$; and 3) greater than $0.7 \cdot h_m$. The third column of Fig. 4 shows masks of different discretized height ranges, and the fourth column shows feature maps of the MHE network. We can see a very high correlation between them.

Different from the semantic selectivity, it can be seen that the regions of buildings and trees are both presented in the activation maps for the medium height range (about 3–15 m). This indicates that the height range selectivity is different from the semantic selectivity. There are two different patterns in the learned deep representations. Furthermore, it also explains why it does not work by using the height threshold for semantic segmentation. We will provide a more detailed analysis in the experiment section.

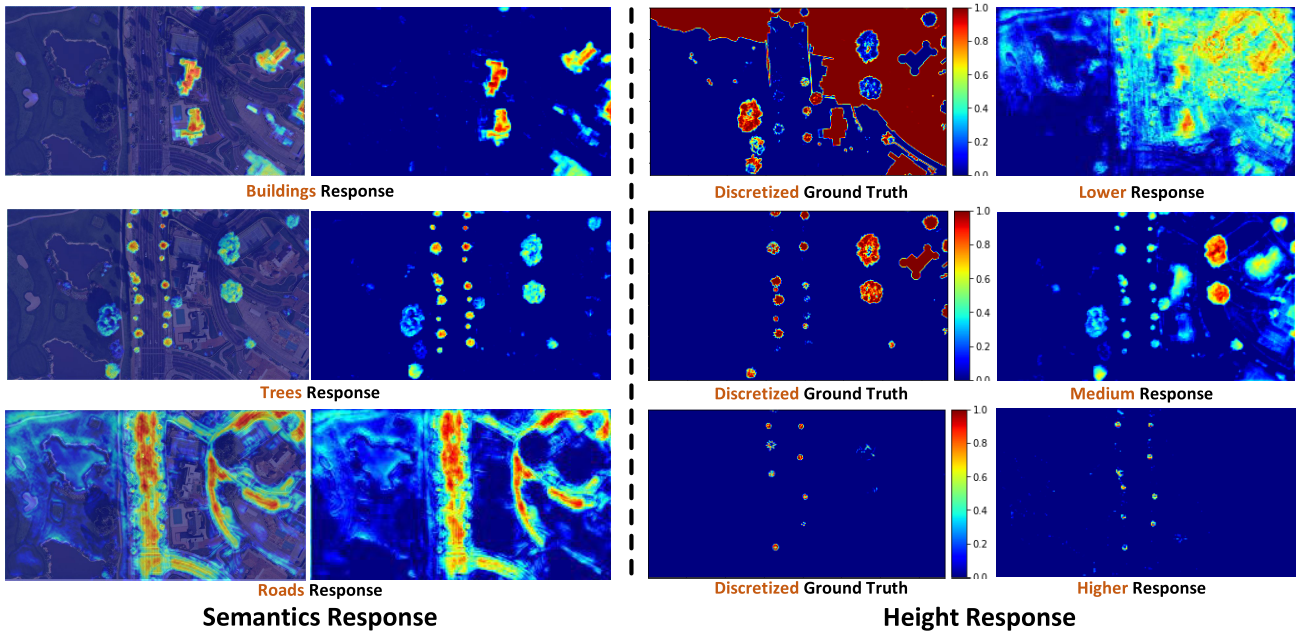


Fig. 4. (Left) Visualization of the high correlation between *semantic maps* and feature maps of MHE networks. (Right) Visualization of the high correlation between *height ranges* and feature maps of MHE networks.

B. MHE for Self-Supervised Pretraining

Our methodology of employing MHE as a self-supervised pretraining is grounded on the pivotal discovery that MHE models can learn semantic representations, distinguishing between various semantic classes such as buildings and trees, purely from the height estimation task. Leveraging MHE for pretraining equips the model with pixel-level semantic insights. As depicted in Fig. 2, we illustrate a framework that transfers the pretrained weights from the MHE task directly to semantic segmentation tasks. This transfer is facilitated through a simple yet effective process:

1) Model Training With MHE: Given an input image $\mathbf{X}$ and its corresponding height ground truth $\mathbf{H}$, we train the deep network $\mathbf{H} = \mathcal{F}_d(\mathcal{F}_e(\mathbf{X}))$ using the mean squared error (MSE) loss, $L = \text{MSE}(\mathbf{H}, \mathbf{H}')$, where $\mathbf{H}'$ is the ground truth height. Here, $\mathcal{F}_d$ represents the height decoder, and $\mathcal{F}_e$ is the backbone encoder of the network. This training step allows the model to learn to predict height for each pixel.

2) Transfer Backbone to Semantic Segmentation: After MHE pretraining, the encoder $\mathcal{F}_e$ used for extracting features relevant to height estimation, is repurposed for semantic segmentation tasks. Given the input image $\mathbf{X}$, the goal is to produce a segmentation map $\mathbf{Y}$, where $\mathbf{Y} = \mathcal{F}_s(\mathcal{F}_e(\mathbf{X}))$ and $\mathcal{F}_s$ denotes the segmentation head. This step leverages the semantic and geometric understanding developed during the MHE pretraining, enhancing the model's performance on semantic segmentation tasks.

C. Disentangled Latent Transformer

To better leverage the semantic selectivity, in this study, we propose to design a DLT module to explicitly model the representations for different semantic classes and height ranges. We find that different groups of activations are selective to different semantic objects. However, there is high redundancy and noise in the deep activation maps within each group. Hence, we propose the DLT module to learn meaningful and robust representations for height estimation. We first cluster these neurons into different semantic groups. Formally, given a pretrained Transformer-based deep network with n neurons at the final layer, the weights $\{\boldsymbol{W_i} \in \mathbb{R}^{n_o \times k \times k}, i = 1, \ldots, n\}$ of the final layer are clustered into K groups using k-means. Note that for each pretrained deep model, we only need to cluster the neurons once and do not need to cluster the feature maps at each inference. Based on the predicted cluster indices, we can further obtain K groups of feature maps $\{\boldsymbol{F_{ci}} \in \mathbb{R}^{n_i \times H \times W}, i = 1, \ldots, K\}$. Each semantic group contains n_i feature maps, and $\sum_{i=1}^{K} n_i = n$.

Feature maps in the same semantic group have similar semantic responses with slight differences. To improve the reliability, we propose to treat the feature maps within the same semantic group as noisy observations and model the true semantic response $\boldsymbol{F_v} \in \mathbb{R}^{K \times H \times W}$ as a latent variable remaining to be estimated. In practice, for each semantic class, we aim to sample the latent semantic responses from a posterior distribution $p(\boldsymbol{F_v} \mid x)$ to explicitly model the noise in a Bayesian probabilistic framework. As the true posterior is intractable to obtain, we adopt the variational inference [41] to approximate $p(\boldsymbol{F_v} \mid x)$ with a simple distribution $q_x(\boldsymbol{F_v})$. Thus, the optimization goal is to minimize the Kullback–Leibler (KL) divergence $\mathcal{L}_{\text{KL}} = \text{KL}(q(\boldsymbol{F_v}) \,||\, p(\boldsymbol{F_v}|x))$. As pointed out by [41], maximizing the likelihood distribution is equivalent to maximizing the variational evidence lower bound (ELBO), which can be defined as

$$\begin{aligned}\mathcal{L}_{\text{KL}} &= \sum_{\boldsymbol{F_v}} q(\boldsymbol{F_v} \mid x) \log\left(\frac{p(\boldsymbol{F_v}, x)}{q(\boldsymbol{F_v} \mid x)}\right) \\ &= \sum_{\boldsymbol{F_v}} q_x(\boldsymbol{F_v}) \log\left(\frac{p(\boldsymbol{F_v})}{q_x(\boldsymbol{F_v})}\right) + \sum_{\boldsymbol{F_v}} q_x(\boldsymbol{F_v}) \log(p(x \mid \boldsymbol{F_v})) \\ &= \mathbb{E}_{q_x(\boldsymbol{F_v})}[\log\ p(\boldsymbol{F_c}|\boldsymbol{F_v})] - \text{KL}(q_x(\boldsymbol{F_v})||p(\boldsymbol{F_v})). \end{aligned} \tag{2}$$

In practice, the first expected log-likelihood term is used to minimize the feature reconstruction loss. Namely, we aim to

reconstruct the original feature maps using the estimated latent variables $\boldsymbol{F_v}$. This can also be viewed as a feature distillation process. Then, the following KL divergence term is used to minimize the distance between $q_x(\boldsymbol{F_v})$ and the true posterior $p(\boldsymbol{F_v} \mid x)$.

To train the whole model in an end-to-end manner using the gradient descent optimizer, we need to compute the gradients with respect to the parameters of distributions. For the sake of simplicity, we assume that the latent semantic responses $\boldsymbol{F_v}$ follow the isotropic Gaussian distribution with mean $\mu_i \in \mathbb{R}^{H,W}$ and variance ${\sigma_i}^2 \in \mathbb{R}^{H,W}$. However, the gradients of parameters $\{\mu_i, {\sigma_i}^2\}$ of the approximate distribution, which is predicted by deep networks, cannot be computed directly. Thus, we utilize the reparameterization trick to rewrite $q_x(\boldsymbol{F_v})$ as $\boldsymbol{F_{vi}} = \mu_i + \sigma_i \varepsilon, i = 1, \ldots, K$, where ε is the standard normal distribution, i.e., $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. With the reparameterization trick, the final loss function can be rewritten as

$$\mathcal{L}(x^{(i)}) = \frac{N}{M}\sum_{i=1}^{M}\left(\frac{1}{L}\sum_{l=1}^{L}\log\ p(h^{(i)}|x^{(i)}, \mu) + \sum \log\ p\left({\boldsymbol{F_c}}^{(i)}|x^{(i)}, \mu + \sigma\varepsilon\right) - \mathrm{KL}(q_x(\boldsymbol{F_v}|x^{(i)})||p(\boldsymbol{F_v}))\right) \quad (3)$$

where M is the size of mini-batch, and N is the total number of samples. L is the number of sampling processes for Monte Carlo estimation. h_i is the ground truth height map for training the height estimation models.

For the prior distribution, we assume that the latent semantic responses follow isotropic Gaussian distribution with mean $\boldsymbol{\mu_i} \in \mathbb{R}^{\boldsymbol{H}\times\boldsymbol{W}}$ and variance $\boldsymbol{\sigma_i}^{\mathbf{2}} \in \mathbb{R}^{\boldsymbol{H}\times\boldsymbol{W}}$. In this work, we use the mean of $\boldsymbol{F_c}$ as the prior mean of $\boldsymbol{F_v}$ and set the variance to be $\boldsymbol{I}$. Then, the latent semantic responses $\boldsymbol{F v_i}$ can be defined as

$$\begin{aligned} \boldsymbol{F}_{\boldsymbol{v}_i} &\sim \mathcal{N}(\mu_p i, \sigma_p i^2) \\ \mu_p i &= \frac{1}{ni}\sum \boldsymbol{F_{ci}}, \quad \sigma_p i = \boldsymbol{I}. \end{aligned} \quad (4)$$

With the defined notations, the KL divergence loss in (3) for distribution approximation can be computed analytically as

$$\mathrm{KL}(q_x(\boldsymbol{F_v})||p(\boldsymbol{F_v})) = -\frac{1}{2}\sum\left[1 + \log\sigma_i^2 - \sigma_i^2 - \mu_i - {\mu_p i}^2\right]. \quad (5)$$

After the training stage, the predicted mean of $\boldsymbol{F_v}$ can be used to generate the semantic segmentation maps for a specific class. In this work, we simply use OTSU [42] algorithm to binarize $\boldsymbol{F_v}$ into segmentation maps. Note that the final segmentation results are obtained from an MHE model without the need for any segmentation annotations.

D. WDC Dataset

The newly introduced WDC dataset, crafted utilizing open data from Open Data DC,[1] encompasses the entirety of WDC, USA. This dataset collects aerial orthophotos captured in 2021 with a normalized nDSM derived from LiDAR data collected in 2020. It features a detailed semantic map that classifies the landscape into five categories: ground, vegetation, buildings, water, and roads. To ensure accuracy and detail in the semantic maps, building footprints, and road layouts were meticulously delineated from shapefiles, while the classification of other categories leveraged the labels present in LiDAR point clouds.

The WDC dataset contains 2159 high-resolution aerial orthophotos, each meticulously paired with a corresponding pixel-wise semantic map and a height map, ensuring a rich dataset for comprehensive analysis. Each image patch, with a resolution of 1024 × 1024 pixels, captures a snapshot of the diverse landscapes encompassed within the dataset, including urban centers, rural expanses, and forested areas. This diversity not only provides a wide array of scenes for analysis but also presents a challenging dataset that encompasses a vast spectrum of classes and height variations. Fig. 5 offers a glimpse into the dataset, showcasing examples that highlight its variety and the detailed nature of its semantic and height maps. The WDC dataset with its broad coverage and high-resolution imagery, stands as a significant resource for advancing research in developing and evaluating models for semantic segmentation and height estimation.

IV. Experiments

In this section, we evaluate the proposed multilevel interpretation framework on three different datasets.

A. Datasets

Three datasets are used in this work for the model evaluation. The first one is a real-world dataset from IEEE GRSS, Data Fusion Contest 2019 (DFC 2019) [43]. DFC 2019 contains 2786 images with both the semantic segmentation label and nDSM. Five semantic classes are included in DFC 2019: buildings, elevated roads and bridges, high vegetation, ground, and water. The second one is a synthetic dataset, called GTAH,[2] which is constructed using the game engine GTA V [44]. GTAH contains 85 881 images with the corresponding height maps. Finally, we also evaluate the proposed method on our newly constructed WDC dataset. These three datasets are all used for evaluating the height estimation performance. As there is no semantic segmentation annotation in GTAH, only DFC 2019 and the WDC dataset are utilized for evaluating the semantic segmentation tasks.

B. Implementation Details

We implement the proposed method using PyTorch [45] and MMSegmentation [46]. On the GTAH dataset [3], we pretrain the Swin Transformer model with 100 epochs and fine-tune the proposed DLT model for 20 epochs. For the U-Net model with ResNet-34 backbone, the code[3] from [9] is used. To optimize the Transformer-based models, AdamW [47] is used with an initial learning rate of 6e-5 for pretraining the deep models. For CNN-based models, Adam [48] is used as the optimizer with an initial learning rate of 1e-4. For all the experiments, we set the batch size to four. For semantic segmentation tasks in the

[1] https://opendata.dc.gov/

[2] https://thebenchmarkh.github.io/

[3] https://github.com/pubgeo/monocular-geocentric-pose

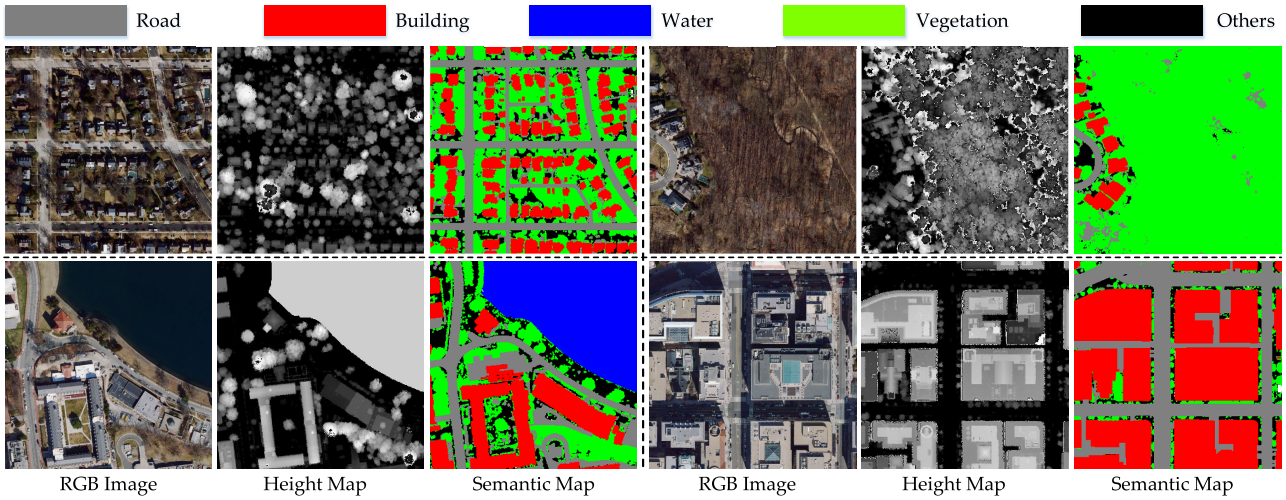


Fig. 5. Visualization of the proposed WDC dataset. Patches of rural, forest, water, and urban are shown as examples. For each patch, the orthophoto, nDSM (height map), and semantic map are displayed.

TABLE I

CLASS-SELECTIVITY AND RESPONSE OF DIFFERENT NEURONS OF SWIN TRANSFORMER ON THE DFC 2019 DATASET

Category	Ground	High Vege.	Building	Water	Selectivity
Neuron #87	0.0833	0.7083	0.0485	0.1597	**0.8893**
Neuron #20	0.0433	0.2051	0.7241	0.0273	**0.8359**
Neuron #255	0.6724	0.1473	0.0606	0.1194	**0.8302**

TABLE II

CLASS-SELECTIVITY AND RESPONSE OF DIFFERENT NEURONS OF UNET (RESNET-34) ON THE DFC 2019 DATASET

Category	Ground	High Vege.	Building	Water	Selectivity
Neuron #17	0.7801	0.1826	0.0069	0.0302	**0.8248**
Neuron #14	0.1419	0.4519	0.2801	0.1259	0.5635
Neuron #16	0.1512	0.2738	0.3993	0.1756	0.5092

transfer learning setting, we train the models for 30 epochs using the AdamW optimizer. The batch size is 16 and the input image size is 512.

C. Network Dissection Analysis and Comparison

To offer a detailed quantitative analysis supporting our findings across various networks, we thereby provide quantitative verification in this section. Specifically, we compute the class-selectivity scores for two models and present the results in Tables I and II. The findings reveal that neurons within both networks exhibit selectivity toward distinct semantic classes. The data presented in the final column of both tables clearly illustrate that the Transformer-based MHE model demonstrates enhanced class selectivity compared to the UNet-based model. This enhanced selectivity can be attributed to the capability of Transformer-based models to learn more sophisticated representations that effectively encapsulate the relationship between height and semantic information. Notably, the semantic classes of tree, ground, and building are observed to be more sensitive to some neurons. This observation aligns with the expectation that these semantic categories can be easily distinguished by height.

D. Height Estimation Results

In this section, we evaluate and compare the performances of our proposed DLT model with previous state-of-the-art (SOTA) methods on the height estimation task. To evaluate the performance of the proposed methods on MHE, we follow the previous work [3] and use four metrics for performance comparisons including mean absolute error (MAE), Root MSE (RMSE), Scale-Invariant RMSE (SI-RMSE), and multiscale gradient error (MSG). MAE is defined as $\text{MAE} = 1/n * \sum |y_i - \hat{y}_i|$, which is used to measure the mean absolute difference between the predicted values and the reference values. RMSE is defined as $\text{RMSE} = (\Sigma(y_i - \hat{y}_i)^2/n)^{1/2}$, which is a commonly used measurement for regression task. The multiscale gradient matching error, MSGE, can be formulated as $\text{MSGE} = (1/M)\sum_{k=1}^{K}\sum_{i=1}^{M}(|\nabla_x R_i^k| + |\nabla_y R_i^k|)$. To care more about the relative relations in the height maps, we also adopt the SI-RMSE and MSGE as metrics for model evaluation. SI-RMSE is defined as $\text{SI-RMSE} = (1/n)\sum_i R_i^2 - (1/n^2)(\sum_i R_i)^2$.

We use the intersection over union (IoU) for different semantic classes as the evaluation metrics for performance comparison on semantic segmentation tasks.

1) Comparisons on the GTAH Dataset: Using the DLT module, we can make each neuron in the last layer to be only sensitive to a specific semantic class or height range. With this neuron compression process, we can obtain a more compact deep neural network with higher interpretability. As presented in Table III, "SwinT (16 Neurons)" denotes the model where the output channels (the number of channels) of the last layer are reduced from 512 to 16.

However, reducing the number of neurons from 512 to 16 significantly limits the model capacity, and also affects the final height estimation performance. Therefore, the performance of model "SwinT (16 Neurons)" is lower than that of the original model (512 Neurons) on the GTAH dataset. To cope with this issue, we propose to distill the feature maps from the original model to the compressed one by encouraging the high feature similarity between them. The DLT model better captures the noise in the features by explicitly modeling

TABLE III

HEIGHT ESTIMATION RESULTS OF DIFFERENT MODELS ON THE GTAH DATASET. THE BEST AND SECOND-BEST RESULTS ARE IN BLUE AND GREEN

Methods	Height Estimation			
	MAE ↓	RMSE ↓	SI-RMSE ↓	MSGE ↓
U-Net (ResNet34) [49]	4.86	6.73	39.51	3.36
U-Net (ResNet101) [49]	3.22	5.81	29.76	2.97
HRNet-FCN [50]	3.14	5.02	24.84	2.61
Swin Transformer [34]	2.97	4.63	21.16	2.40
Swin (16 Neurons)	3.09	4.75	22.04	2.53
Swin (16 Neurons)+DLT	2.97	4.73	21.61	2.52
Twins [51]	2.31	4.12	18.54	2.15
Twins (16 Neurons)	2.64	4.57	20.18	2.46
Twins (16 Neurons)+DLT	2.49	4.36	20.04	2.27

TABLE IV

HEIGHT ESTIMATION RESULTS OF DIFFERENT METHODS ON THE DFC 2019 DATASET. THE BEST AND SECOND-BEST RESULTS ARE IN BLUE AND GREEN

Method	Height Estimation			
	MAE ↓	RMSE ↓	SI-RMSE ↓	MSGE ↓
U-Net (ResNet-34) [49]	1.43	2.42	7.11	4.02
U-Net (ResNet-101) [49]	1.32	2.31	6.42	3.39
HRNet-FCN [50]	1.30	2.22	6.39	3.22
Swin Transformer [34]	1.24	2.11	5.33	3.03
SwinT (16 Neurons)	1.29	2.21	5.67	2.98
SwinT (16 Neurons)+DLT	1.27	2.18	5.59	3.02
Twins [51]	1.13	1.93	5.08	2.82
Twins (16 Neurons)	1.25	2.15	5.41	2.94
Twins (16 Neurons)+DLT	1.19	2.02	5.28	2.89

the features as latent random variables, thus obtaining similar results to the original model.

To evaluate the applicability of DLT across different SOTA network architectures, we report the results using the Twins Transformer. As delineated in Table III, the Twins-SVT-L model, integrated with UperNet and pretrained on ImageNet-1k, consistently outperforms other models, attributable to its powerful backbone. Incorporating DLT with the Twins backbone demonstrates the generalizability of our findings regarding semantic selectivity derived from MHE pretrained weights across diverse network architectures.

2) Comparisons on the DFC 2019 Dataset: From the results in Table IV, it can be seen that our proposed "SwinT (16 Neurons) + DLT" with only 16 neurons can achieve competitive height estimation performance compared with the original model. This shows the effectiveness of the DLT module on the real-world dataset. From the results, we can also observe that Transformer-based models can obtain better performance when compared with CNN-based models. On the DFC 2019 dataset with the Twins-SVT-L [51] backbone, we observe a consistent trend where employing the DLT module enhances the MHE performance and efficiency. Utilizing the DLT module enables the model to perform unsupervised semantic segmentation, with only a marginal decrease in MHE performance.

3) Comparisons on the WDC Dataset: On the proposed WDC dataset, we can also observe that Transformer-based models have advantages in terms of height estimation performance from Table V. The proposed "SwinT (16 Neurons) + DLT" obtains similar performance compared with the "Swin Transformer" model. It shows that the compressed DLT model can improve the interpretability and reduce the computational complexity of the original Transformer model, with almost no loss of accuracy. On the proposed WDC dataset, models using Twins-SVT-L backbone again outperform others. Utilizing DLT can improve the model efficiency (using only 16 neurons) and interpretability with only a marginal decrease in MHE performance.

TABLE V

HEIGHT ESTIMATION RESULTS OF DIFFERENT METHODS ON THE PROPOSED WDC DATASET. THE BEST AND SECOND-BEST RESULTS ARE IN BLUE AND GREEN

Method	Height Estimation			
	MAE ↓	RMSE ↓	SI-RMSE ↓	MSGE ↓
U-Net (ResNet-34) [49]	4.86	6.27	39.67	6.92
U-Net (ResNet-101) [49]	4.48	5.83	35.05	6.25
HRNet-FCN [50]	4.27	6.04	35.60	6.64
Swin Transformer [34]	4.15	5.40	31.96	5.86
SwinT (16 Neurons)	4.41	5.73	35.61	6.14
SwinT (16 Neurons)+DLT	4.35	5.71	33.89	6.09
Twins [51]	3.71	5.05	27.63	5.46
Twins (16 Neurons)	4.11	5.42	31.44	5.93
Twins (16 Neurons)+DLT	3.82	5.30	30.96	5.61

E. MHE Pretraining for Semantic Segmentation

This section presents the results of leveraging pretrained weights from MHE to enhance downstream semantic segmentation tasks on the WDC and the DFC 2019 dataset. We use the Swin-Base Transformer as the backbone. Our comparative analysis encompasses four distinct transfer learning settings: 1) training from scratch; 2) utilizing ImageNet pretrained weights; 3) applying MHE pretrained weights with the backbone fixed while only fine-tuning the segmentation head; and 4) employing MHE pretrained weights for a full fine-tuning of the segmentation model. The experimental results are presented in Tables VI and VII.

The evaluation of these methodologies yields insightful findings. Primarily, using ImageNet pretrained weights can boost segmentation performance compared with training from scratch, underscoring the importance of transfer learning for RS segmentation tasks. More intriguingly, the utilization of MHE pretrained weights not only outperforms the training-from-scratch model but also showcases substantial performance improvement compared with the model using ImageNet pretrained weights. This enhancement is more notable when the model is fully fine-tuned in downstream segmentation tasks, indicating the effectiveness of MHE task-derived representations for pixel-level RS tasks.

On the WDC dataset, we can observe clear improvements in segmentation performance for the tree, building, road, and ground categories using the pretrained weights from MHE. This makes sense as these categories are more related to the height attribute. The superior performance of models utilizing MHE pretrained weights substantiates the hypothesis that the MHE task can learn rich semantic representations from 3-D elevation data. These representations, in turn, provide a

TABLE VI
SEGMENTATION RESULTS OF DIFFERENT METHODS ON THE PROPOSED WDC DATASET. THE BEST RESULTS ARE PRESENTED IN BOLD

Methods	mIoU					Acc				
Categories	ground	vegetation	building	water	road	ground	vegetation	building	water	road
Scratch	48.54	66.63	73.68	61.05	62.3	61.86	83.39	84.08	71.23	76.11
ImageNet	51.64	68.83	77.64	**69.91**	66.59	67.67	81.67	91.59	**87.46**	77.24
MHE+SegFinetune	55.06	71.03	81.12	66.47	70.02	70.43	83.37	91.88	76.41	81.00
MHE+FullFinetune	**56.60**	**71.89**	**82.87**	67.60	**71.23**	**71.15**	**85.94**	**92.27**	79.91	**82.90**

TABLE VII
SEGMENTATION RESULTS OF DIFFERENT METHODS ON THE DFC 2019 DATASET. THE BEST RESULTS ARE PRESENTED IN BOLD

Methods	mIoU					Acc				
Categories	ground	tree	building	water	bridge	ground	tree	building	water	bridge
Scratch	82.44	48.53	50.03	71.09	37.62	91.97	60.88	69.96	72.56	40.40
ImageNet	84.74	52.01	56.33	83.71	67.61	92.98	63.50	75.11	84.92	84.86
MHE+SegFinetune	86.67	54.70	63.64	83.70	69.95	94.29	66.32	79.20	**86.01**	80.99
MHE+FullFinetune	**86.91**	**55.61**	**63.67**	**84.34**	**73.21**	**94.31**	**67.63**	**80.03**	**87.16**	**86.58**

better understanding of the semantic intricacies inherent in RS imagery. Similarly, on the DFC 2019 dataset, the experimental results highlight the benefits of employing MHE pretrained weights by a marked improvement in segmentation accuracy. The simplicity and effectiveness of MHE pretraining signify a promising direction for future research on bridging the gap between height estimation and semantic segmentation. We offer a novel pathway to integrate elevation data in RS models and showcase the untapped potential of MHE for training RS foundation models on pixel-level dense prediction tasks.

F. Unsupervised Segmentation Results

As we use the height data as supervision, the segmentation task can also be viewed as a *unsupervised segmentation* task. We mainly focus on the segmentation performance of *Building* and *Tree* classes that are important for urban planning, climate change, and disaster monitoring. Annotating pixel-level semantic labels is expensive for global-scale Earth observation applications, while our proposed DLT model can obtain the pixel-level semantic segmentation maps using only the height maps as supervision. This has great potential for large-scale applications because 3-D models for many cities have been provided.

We first compare with a simple baseline, which uses thresholds to split the predicted height map into different semantic classes. The accuracy is poor *since many buildings and trees are in the same height range*. In the following part, we will explain why simple threshold-based methods fail. According to [55], the minimum house building is about 3 m. Thus, in this study, we set the height range to $3 \sim h_{\max}$ meters, where $h_{\max}$ is the maximum height in the given image. For the tree canopy height range, we set it to 10∼25 m according to the statistics about the tree height [56].

Some visualization examples of the segmentation results using a threshold-based method are shown in Fig. 6. It can be seen that buildings and trees are heavily entangled due to the large overlap of height ranges. On one hand, the height values are quite useful features to distinguish different ground objects, as height is an inherent attribute of these objects. On the other hand, how to make better use of these features is still an open problem, as the simple threshold-based method fails in most cases.

TABLE VIII
UNSUPERVISED SEMANTIC SEGMENTATION RESULTS ON THE DFC 2019 DATASET. THE BEST RESULTS ARE PRESENTED IN BOLD

Method	Semantic Segmentation		
	Building(IoU)	Tree(IoU)	Mean
Threshold	0.126	0.237	0.182
PiCIE [52]	0.150	0.280	0.215
IIC [53]	0.157	0.275	0.216
MaskContrast* [54]	0	0.0001	0.0001
RGBH, PiCIE [52]	0.167	0.294	0.231
RGBH, IIC [53]	0.173	0.291	0.232
RGBH, MaskContrast* [54]	0.026	0.003	0.015
U-Net (16 Neurons) [49]	0.092	0.077	0.0845
HRNet-FCN [50]	0.146	0.277	0.212
SwinT (16 Neurons)	0.269	0.348	0.309
SwinT (16 Neurons)+DLT	**0.306**	**0.389**	**0.348**

Then, we compare our results with three unsupervised semantic segmentation networks, including MaskContrast [54], PiCIE [52], and IIC [53]. MaskContrast uses saliency as the prior knowledge for contrastive learning and gains SOTA performance for natural images. However, RS images are quite different from natural images in terms of contexts and scales. In our experiments, we find that the saliency-based method fails to distinguish foreground and background on DFC dataset. PiCIE considers the photometric and geometric invariance of pixel embeddings. It also suffers from the difficulties caused by the differences between natural and RS images. Although IoU scores are

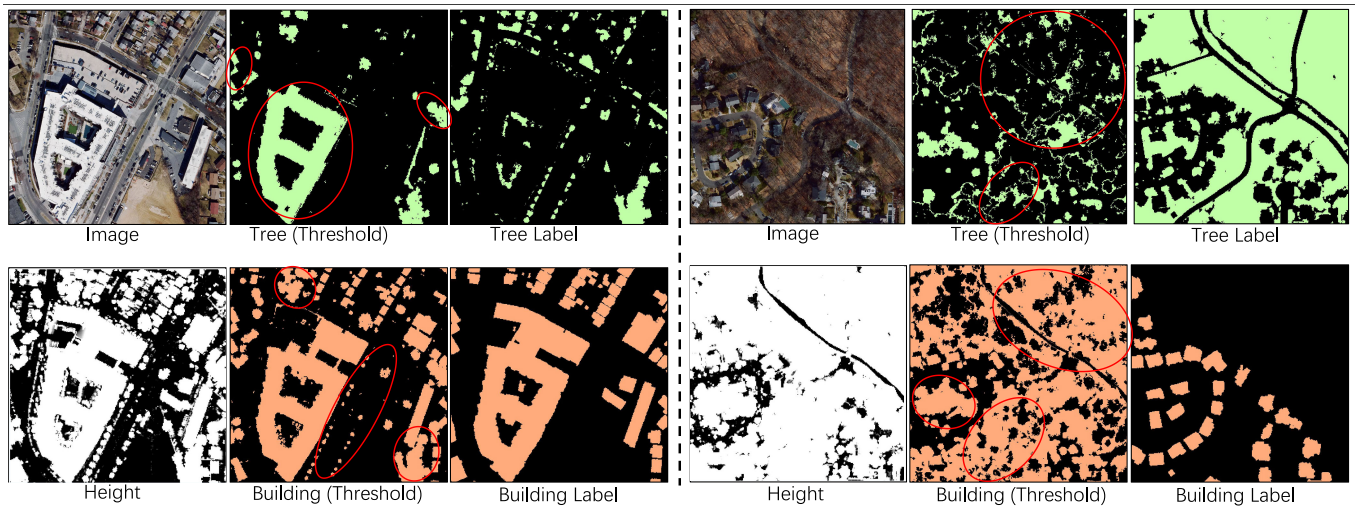


Fig. 6. Visualization of the semantic segmentation results of using a simple threshold baseline. Both the building and tree segmentation results are displayed. From the figure, it can be seen that buildings and trees are heavily entangled due to the large overlap of height ranges.

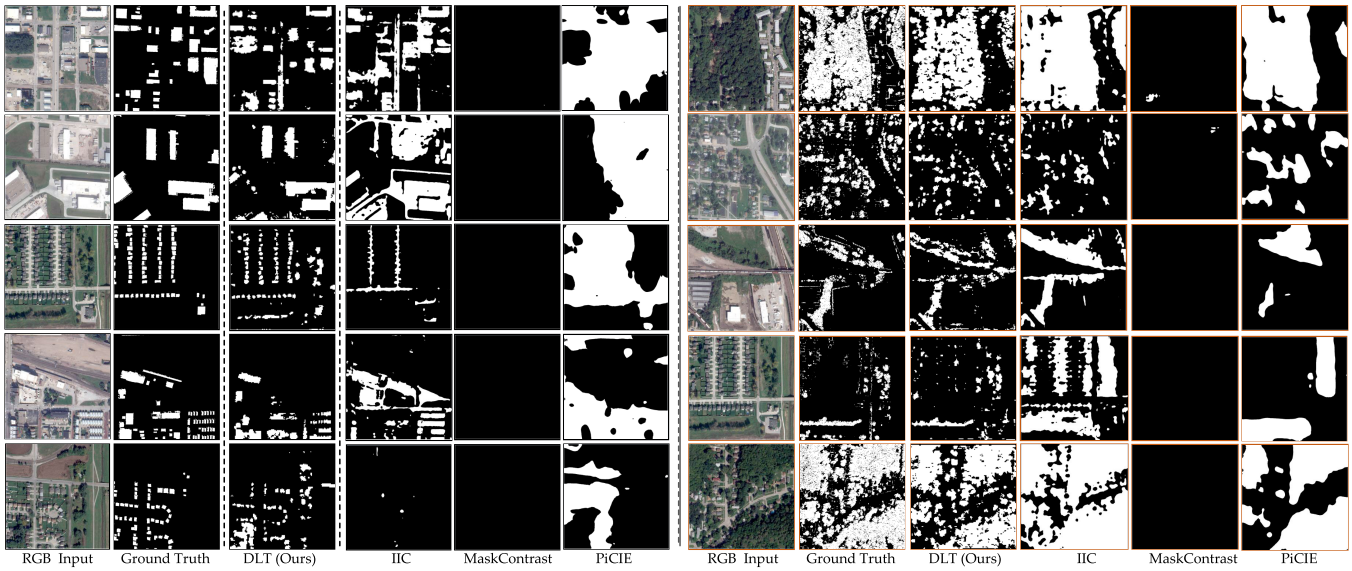


Fig. 7. Unsupervised semantic segmentation results of the proposed "SwinT (16 Neurons) + DLT" on the DFC 2019 dataset. (Left) Segmentation maps of Neuron no.3 (building) and (right) Neuron no. 7 (tree) are shown in this figure. Compared with its counterparts, our DLT can obtain better segmentation results (best viewed by zooming in).

high, the predictions do not capture the layouts of input images.

Toward a fair comparison, we also use the height data in addition to RGB images as the input of the unsupervised segmentation methods. Although the results are improved, they are still much lower than the proposed method. *Note that the proposed DLT model does not need the height input during the test phase.* The results in Table VIII show that DLT can outperform existing unsupervised semantic segmentation methods clearly by a large margin. For example, compared with IIC, DLT can increase the IoU metric from 0.157 to 0.306. As shown in Table IX, on the WDC Dataset, we can also observe that compared with MaskContrast [54], PiCIE [52], and IIC [53], the proposed DLT model can achieve much better segmentation performance. The aforementioned experimental results have demonstrated the effectiveness of our proposed DLT model on both height estimation and unsupervised segmentation tasks. In Figs. 7 and 8, we visualize some qualitative examples of our DLT method and its counterparts for a clear comparison.

TABLE IX
UNSUPERVISED SEMANTIC SEGMENTATION RESULTS ON THE PROPOSED WDC DATASET. THE BEST RESULTS ARE PRESENTED IN BOLD

Method	Semantic Segmentation		
	Building(IoU)	Tree(IoU)	Mean
Threshold	0.178	0.213	0.196
PiCIE [52]	0.278	0.221	0.250
IIC [53]	0.201	0.289	0.245
MaskContrast* [54]	0.002	0.368	0.185
RGBH, PiCIE [52]	0.292	0.236	0.264
RGBH, IIC [53]	0.213	0.302	0.258
RGBH, MaskContrast* [54]	0.015	0.372	0.194
U-Net (16 Neurons) [49]	0.113	0.127	0.120
HRNet-FCN [50]	0.153	0.274	0.214
SwinT (16 Neurons)	0.336	0.389	0.363
SwinT (16 Neurons)+DLT	**0.365**	**0.412**	**0.389**

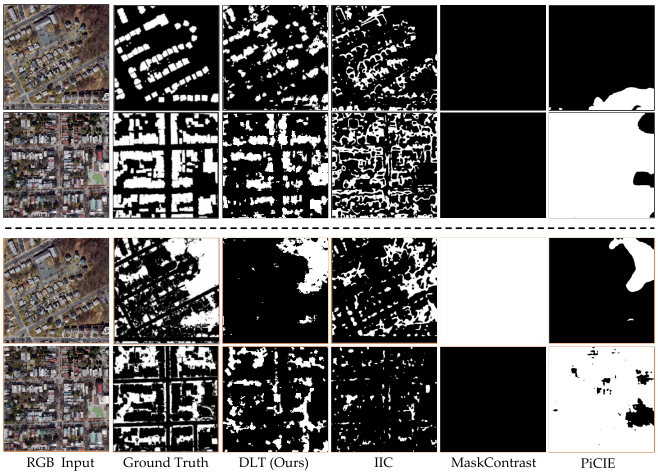


Fig. 8. Unsupervised semantic segmentation results of the proposed "SwinT (16 Neurons) + DLT" on the WDC dataset. WDC is a challenging dataset for unsupervised segmentation due to the season change. Compared with its counterparts, our DLT can obtain better segmentation results (best viewed by zooming in).

V. Conclusion

This article explores how deep networks predict height from single images. To achieve this, we explore the deep MHE models by a neuron-level network dissection and find that the internal deep representations of the MHE model have a high semantic and height range selectivity. Motivated by this finding, we further make three contributions: 1) we design a self-supervised pretraining framework based on the MHE task; 2) we introduced a DLT network toward a compact and explainable deep model for joint semantic segmentation and height estimation; and 3) we construct a new benchmark dataset for research on both semantic segmentation and height estimation tasks. The proposed model enjoys a free lunch, i.e., semantic segmentation, by learning only from the height data. The effectiveness of the proposed method is demonstrated on three datasets including height estimation and semantic segmentation tasks. In summary, the findings in this work provide novel insights for designing new self-supervised pretraining, semantic segmentation, and height estimation models.

References

[1] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[2] Z. Xiong, F. Zhang, Y. Wang, Y. Shi, and X. Xiang Zhu, "EarthNets: Empowering AI in Earth observation," 2022, *arXiv:2210.04936*.

[3] Z. Xiong, W. Huang, J. Hu, and X. X. Zhu, "THE benchmark: Transferable representation learning for monocular height estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 10251159.

[4] B. Le Saux, N. Yokoya, R. Hansch, M. Brown, and G. Hager, "2019 data fusion contest [technical committees]," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 1, pp. 103–105, Mar. 2019.

[5] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2, 2014, pp. 2366–2374.

[6] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2002–2011.

[7] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6647–6655.

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

[9] 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.

[10] 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.

[11] Y. Sun, L. Mou, Y. Wang, S. Montazeri, and X. X. Zhu, "Large-scale building height retrieval from single SAR imagery based on bounding box regression networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 79–95, Feb. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271621003221

[12] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. 8, pp. 42200–42216, 2020.

[13] Q. Zhang, Y. N. Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8827–8836.

[14] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, "This looks like that: Deep learning for interpretable image recognition," 2018, *arXiv:1806.10574*.

[15] T. Wu and X. Song, "Towards interpretable object detection by unfolding latent structures," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6032–6042.

[16] 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.

[17] J. Hu, Y. Zhang, and T. Okatani, "Visualization of convolutional neural networks for monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3868–3877.

[18] Z. You, Y.-H. Tsai, W.-C. Chiu, and G. Li, "Towards interpretable deep networks for monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12859–12868.

[19] X. Xiang Zhu et al., "On the foundations of Earth and climate foundation models," 2024, *arXiv:2405.04285*.

[20] Z. Xiong et al., "Neural plasticity-inspired foundation model for observing the Earth crossing modalities," 2024, *arXiv:2403.15356*.

[21] A. Fuller, K. Millard, and J. Green, "CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 5506–5538.

[22] Y. Wang, C. M. Albrecht, N. A. A. Braham, C. Liu, Z. Xiong, and X. X. Zhu, "DeCUR: Decoupling common & unique representations for multimodal self-supervision," 2023, *arXiv:2309.05300*.

[23] S. Liao and L. Shao, "Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K.: Springer, 2020, pp. 456–474.

[24] H. Liang et al., "Training interpretable convolutional neural networks by differentiating class-specific filters," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 622–638.

[25] I. Obadic, R. Roscher, D. A. B. Oliveira, and X. X. Zhu, "Exploring self-attention for crop-type classification explainability," 2022, *arXiv:2210.13167*.

[26] J. Garcke and R. Roscher, "Explainable machine learning," *Mach. Learn. Knowl. Extraction*, vol. 5, no. 1, pp. 169–170, 2023, doi: 10.3390/make5010010.

[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[28] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Zürich, Switzerland: Springer, 2014, pp. 818–833.

[29] J. Gu and C. Dong, "Interpreting super-resolution networks with local attribution maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9199–9208.

[30] 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.

[31] 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.

[32] S. Kunwar, "U-Net ensemble for semantic and height estimation using coarse-map initialization," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 4959–4962.

[33] 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.

[34] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[35] S. Chen, Y. Shi, Z. Xiong, and X. X. Zhu, "HTC-DC net: Monocular height estimation from single remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 10294289.

[36] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 9782149.

[37] Y. Cong et al., "SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 197–211.

[38] 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.

[39] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," 2023, *arXiv:2311.07113*.

[40] A. S. Morcos, D. G. T. Barrett, N. C. Rabinowitz, and M. Botvinick, "On the importance of single directions for generalization," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[41] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[42] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[43] B. Le Saux, N. Yokoya, R. Hänsch, and M. Brown, 2019, "Data fusion contest 2019 (DFC2019)," *IEEE DataPort*, doi: 10.21227/C6TM-VW12.

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

[45] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8024–8035.

[46] M. Contributors. (2020). *MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark*. [Online]. Available: https://github.com/open-mmlab/mmsegmentation

[47] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[49] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[50] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2020.

[51] X. Chu et al., "Twins: Revisiting the design of spatial attention in vision transformers," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 9355–9366.

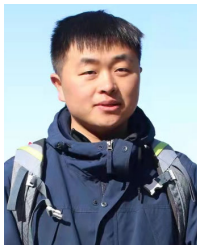
[52] J. Hyun Cho, U. Mall, K. Bala, and B. Hariharan, "PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16789–16799.

[53] X. Ji, A. Vedaldi, and J. Henriques, "Invariant information clustering for unsupervised image classification and segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9865–9874.

[54] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, and L. Van Gool, "Unsupervised semantic segmentation by contrasting object mask proposals," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10032–10042.

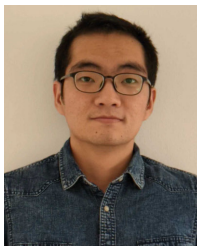
[55] L. Appolloni and D. D'Alessandro, "Housing spaces in nine European countries: A comparison of dimensional requirements," *Int. J. Environ. Res. Public Health*, vol. 18, no. 8, p. 4278, Apr. 2021.

[56] X. Liu et al., "Neural network guided interpolation for mapping canopy height of China's forests by integrating GEDI and ICESat-2 data," *Remote Sens. Environ.*, vol. 269, Feb. 2022, Art. no. 112844.



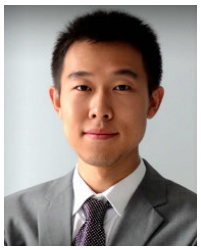
Zhitong Xiong (Member, IEEE) received the Ph.D. degree in computer science and technology from the Northwestern Polytechnical University, Xi'an, China, in 2021.

He is currently a Research Scientist with the Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany. His research interests include computer vision, machine learning, and remote sensing.



Sining Chen received the bachelor's degree in marine science from Xiamen University, Xiamen, China, in 2018, and the master's degree in Earth-oriented space science and technology (ESPACE) from Technical University of Munich (TUM), Munich, Germany, in 2020, where he is currently pursuing the Ph.D. degree with the Chair of Data Science in Earth Observation.

He was a DLR/DAAD Doctoral Research Fellow with the German Aerospace Center (DLR), Remote Sensing Technology Institute, Wessling, Germany, from September 2021 to August 2023. His research interests include deep learning, monocular height estimation, and 3-D building reconstruction.



Yilei Shi (Member, IEEE) received the Dipl.-Ing. degree in mechanical engineering and the Dr.Ing. degree in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2010 and 2019, respectively.

In April and May 2019, he was a Guest Scientist with the Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, U.K. He is currently a Senior Scientist with the Chair of Remote Sensing Technology, Technical University of Munich.

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

She was the Founding Head of the Department "EO Data Science", German Aerospace Center (DLR), Remote Sensing Technology Institute, Wessling, Germany. 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, Los Angeles, CA, USA, in 2009, and 2014-2016, respectively. Since May 2020, she has been the PI and the Director of the International Future AI Laboratory "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond". Since October 2020, she has been a Director of Munich Data Science Institute (MDSI), TUM. From 2019 to 2022, she was a coordinator of Munich Data Science Research School (www.muds.de) and the Head of the Helmholtz Artificial Intelligence – Research Field "Aeronautics, Space and Transport." She is the Chair Professor of data science in Earth observation with TUM. She is currently a Visiting AI Professor with ESA's Phi-lab, Frascati, Italy. Her main research interests include are remote sensing and Earth observation, signal processing, machine learning and data science, with their applications in tackling societal grand challenges, e.g., global urbanization, UN's SDGs, and climate change.

Dr. Zhu has been a member of the Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is a fellow of the Academia Europaea (the Academy of Europe), AAIA, and ELLIS. She serves in the Scientific Advisory Board in several research organizations, among others the German Research Center for Geosciences (GFZ, 2020–2023) and Potsdam Institute for Climate Impact Research (PIK). She served as a Area Editor responsible for special issues of *IEEE Signal Processing Magazine* (2021–2023). She is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and *Pattern Recognition*.