HTC-DC Net: Monocular Height Estimation From Single Remote Sensing Images

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

Abstract— Three-dimensional geoinformation is of great significance for understanding the living environment; however, 3-D perception from remote sensing data, especially on a large scale, is restricted, mainly due to the high costs of 3-D sensors such as light detection and ranging (LiDAR). To tackle this problem, we propose a method for monocular height estimation from optical imagery, which is currently one of the richest sources of remote sensing data. As an ill-posed problem, monocular height estimation requires well-designed networks for enhanced representations to improve the performance. Moreover, the distribution of height values is long-tailed with the low-height pixels, e.g., the background (BG), as the head, and thus, trained networks are usually biased and tend to underestimate building heights. To solve the problems, instead of formalizing the problem as a regression task, we propose HTC-DC Net following the classification–regression paradigm, with the head-tail cut (HTC) and the distribution-based constraints (DCs) as the main contributions. HTC-DC Net is composed of the backbone network as the feature extractor, the HTC-AdaBins module, and the hybrid regression process. The HTC-AdaBins module serves as the classification phase to determine bins adaptive to each input image. It is equipped with a vision transformer (ViT) encoder to incorporate local context with holistic information and involves an HTC to address the long-tailed problem in monocular height estimation for balancing the performances of foreground (FG) and BG pixels. The hybrid regression process does the regression via the smoothing of bins from the classification phase, which is trained via DCs. The proposed network is tested on three datasets of different resolutions, namely ISPRS Vaihingen (0.09 m), Data Fusion Contest 19 (DFC19) (1.3 m), and Global Building Height (GBH) (3 m). The experimental results show the superiority of the proposed network over existing methods by large margins. Extensive ablation studies demonstrate the effectiveness of each design component. The codes and trained models are published at https://github.com/zhu-xlab/HTC-DC-Net.

Index Terms— Adaptive bins, hybrid regression, monocular height estimation, vision transformer (ViT).

I. INTRODUCTION

MONOCULAR height estimation is the process of deriving height information from single remote sensing images. The generated height maps, usually delivered in the form of digital surface models (DSMs) or normalized DSMs (nDSMs), are of great importance for many downstream applications. For example, estimating building heights is essential for 3-D building models [1], [2], [3], which serves as a crucial information basis for urban planning and disaster management. Also, modeling vegetation heights, represented as canopy height models [4], [5], [6], can improve the understanding of biomass and, thus, help the carbon cycle studies on a large scale.

Height information can be retrieved directly with 3-D-aware techniques, e.g., 3-D sensors such as light detection and ranging (LiDAR) [7], [8], [9] and synthetic aperture radar (SAR) [10], [11], or stereo pairs of optical images. However, such techniques are only conditionally applicable for various reasons. Though LiDAR delivers high-quality 3-D measurements, the very high operational costs hamper its use in most cases. SAR has larger coverage than LiDAR, and satisfactory accuracy, however, suffers a lot from the side-looking geometry [12], [13]. In dense urban areas, the need for a stack of SAR images restricts its applicability for 3-D reconstruction in practice [14]. While stereo images are easier to obtain, the compromise between acquisition quality and quantity poses great difficulties [15], [16]. Large-scale applications demand high-quality and comprehensive data, which is not adequately met by either costly aerial imaging acquisitions or low-budget satellite stereo pairs. While the former provides high-quality data, it comes at a significant expense, whereas the latter is typically affected by cloud contamination and long baselines, limiting its usefulness for large-scale applications [17].

In contrast, monocular images, especially those from satellites, are rich in quantity [18], which addresses the deficiencies

of the aforementioned techniques and, thus, can support large-scale applications as well as the corresponding updates. The only problem is how to mine the concealed height information from them. Early works on monocular height estimation focus on the level of instances [19], [20]. Following the physical model of shadow casting, shadow lengths are exploited as the cue for inferring heights. Together with solar parameters, the heights of ground objects can be computed mathematically. However, such methods suffer from overlaps between shadows and objects, especially in dense urban areas or dense forests, as well as the availability of exact solar parameters. Fortunately, a large amount of data and the recent emergence of deep learning methods make it possible to tackle the problem in a data-driven manner. Given that sufficient data could be used for training, models of high performance could be expected.

Monocular height estimation could be inspired by advances in monocular depth estimation [21], which is faced with exactly the same problem as monocular height estimation, that being the ill-posed nature. Namely, multiple height maps with similar height structures could look very similar in the domain of optical images; thus, one specific optical image can correspond to multiple height map predictions that are hard to disentangle. Inspired by the use of vision transformers (ViTs) [22] for enforcement of global consistency in monocular depth estimation, we propose involving a ViT for modeling long-range attention to combat the ill-posed problem. Besides, changes in solution paradigms have occurred in monocular depth estimation. Instead of solving the problem as a regression task, the state-of-the-art solution is to convert the problem into a classification–regression problem. For example, the hybrid regression process [23] is proposed to facilitate the solution for monocular depth estimation. In this article, we demonstrate the feasibility and superiority of applying the classification–regression paradigm for monocular height estimation from remote sensing images.

Different from monocular depth estimation, monocular height estimation also suffers from the long-tailed distribution problem [24]. Specifically, in the physical world, most of the ground objects are of lower height, such as low buildings and vegetation, while high objects are rare, e.g., skyscrapers. When a network is trained with such data from nature, it will be largely biased. Considering that the long-tailed distribution of height values is even more skewed than the worst cases in long-tailed classification, the predictions can include many fatal cases for higher objects, with incredibly large errors.

Different from a simple regression process, the hybrid regression process incorporates a distribution-based approach, which yields a distribution specified by the bin centers from the classification phase and the bin probabilities from the regression phase. Theoretically, the final prediction lies within the bins with the highest probabilities. In practice, to avoid discrete predicted values, the final prediction is computed as the weighted average of the bin centers according to the bin probabilities, equivalent to the expectation value of a distribution. This is based on the assumption that the expectation value of the distribution is close to the value where the probability is the highest. However, the assumptions cannot be guaranteed without any constraints set on the distribution.

To cope with the aforementioned problems, we propose HTC-DC Net, which is equipped with a head-tail cut (HTC) and distribution-based constraints (DCs).

In summary, our contributions are given as follows.

1) We propose a novel architecture for monocular height estimation. We utilize a classification–regression paradigm for HTC-DC Net, which employs the ViT for enhanced representation learning.
2) We propose an HTC to address the extremely long-tailed nature of height values, i.e., to mitigate the side impact of the background (BG) pixels as the majority.
3) We propose using DCs to regularize the bin probabilities used during the regression phase, which are mathematically neat and lead to remarkable improvements.
4) We conduct extensive experiments to showcase the efficacy of the proposed network and comprehensive ablation studies to demonstrate the necessity of each designed component. The proposed network outperforms the existing methods by a large margin.

The remainder of this article is organized as follows. Section II gives an overview of the related works. The proposed method is described in detail in Section III, followed by Section IV describing the experiments and Section V showing the experimental results. Discussions and ablation studies are presented in Section VI. Conclusions are drawn, and further research directions are described in Section VII.

II. RELATED WORKS

A. Monocular Height Estimation

Deep-learning-based monocular height estimation networks can be categorized into pixelwise and instancewise methods, based on their distinct objectives. As the major focus of this article, pixelwise height estimation can be formalized as a deep dense regression task. To tackle the task, encoder–decoder fully convolutional networks (FCNs) are mostly utilized [25]. FCNs for semantic segmentation can be adopted by removing the final classification layer (softmax or sigmoid activation), e.g., SegNet [26], U-Net [27], and Eff-UNet [28]. Besides, many networks are proposed specifically for monocular height estimation. Those networks can be categorized into two types: single-task learning networks and multitask learning networks.

1) Single-Task Learning Networks: Single-task learning networks map the input images to the output height maps. In such networks, feature fusion is usually applied to boost performance. For instance, Mou and Zhu [29] proposed one of the first deep-learning-based methods to estimate height from a single optical image—an encoder–decoder neural network, IM2HEIGHT. Compared to plain FCN, IM2HEIGHT has a skip connection, accounting for low-level feature fusion. It leads to sharper object edges and more details preserved. Amirkolaee and Arefi [30] adopted advanced techniques from the CV community. They used the upsampling block to lighten the computation burden and utilized multilevel feature fusion to combat the blurring effect during inference [31]. Besides, they also proposed a postprocessing scheme to enforce the

continuities around the patch edges. Xing et al. [32] proposed PLNet with the feature fusion module—gated feature aggregation module (GFAM)—and a refining module—progressive refine module (PRM).

2) Multitask Learning Networks: Multitask learning networks introduce auxiliary tasks in addition to height predictions, with the expectation that both tasks support each other during training. Usually, based on the assumption that heights and semantics are highly correlated [33], semantic segmentation can be regarded as an auxiliary task for height estimation. For example, Srivastava et al. [34] were the first to showcase the gains the auxiliary semantic segmentation head brings. Carvalho et al. [35] explored earlier separation between heights and semantics and compared different multitask learning strategies. Elhousni et al. [36] used further auxiliary geometric information, the normal vectors, in a two-stage network, where the first-stage results are fed into the second-stage denoising autoencoder for refinement.

As an alternative, monocular height estimation can also be regarded as an image translation task, assuming that the images and heights are backed by the same underlying semantics. In this context, generative adversarial networks (GANs) are used. Ghamisi and Yokoya [37] used a GAN-based network consisting of a generator and a discriminator. The generator applies the style transfer, i.e., takes the input image, and transfers it to the output height map. The discriminator is used to help train the generator to generate realistic height maps. The network is trained with image–height map pairs. Later, Paoletti et al. [38] overcame the problem by introducing the shared latent features. This makes the network generalize better and learn more generic style information. Improvements in the performances are demonstrated by experiments.

Besides, methods for instancewise monocular height estimation, though not the focus of this article, account for cases when doing 3-D perception specifically for some ground objects, e.g., buildings. Under such circumstances, instance segmentation-based networks can be used, where heights are predicted based on the instances as the prior. By this means, the output maps are usually sparse maps with only object pixels filled with objectwise single height values, which, in the context of 3-D building reconstruction, are exactly the LoD-1 (level of details) building models. Such methods are usually based on two-stage instance segmentation networks, e.g., Mask R-CNN [39]. Mahmud et al. [40] modified Mask R-CNN into a multitask network, with a joint prediction of heights, signed distance function, and semantics aggregated into the final output. Chen et al. [41] proposed a network named Mask-Height R-CNN, which is adapted from Mask R-CNN for monocular height estimation by adding a height regression head to the region proposal network (RPN) [42]. Recently, Li et al. [43] proposed a novel type of representation for building instances in 3-D space: 3-D centripetal shift representation. Their proposed network, termed 3DCentripetalNet, learns 3-D centripetal shift representation and building corners, which are further utilized to retrieve building heights.

B. Monocular Depth Estimation

As a highly related task to monocular height estimation, monocular depth estimation has been a long-standing task in the computer vision community [21]. The advances in monocular depth estimation can thus inspire better solutions for monocular height estimation. It has been witnessed that the paradigm of doing monocular depth estimation changes from regression to classification and then to classification–regression—the state-of-the-art solution.

Treated as a regression problem, the depths are predicted directly from the images, which are supervised by the ground truth depth values. There have been many works in this direction, and they are all intuitive; however, their performances are limited [31], [44], [45], [46], [47]. Fu et al. [48] proposed using ordinal regression, which converts the regression problem into a classification problem, which inspired the application of DORN in MHE [49]. Besides, Sun et al. [42] designed a classification network based on the ordinal regression network, however, with an adaptive bin design and a set prediction framework for bin prediction. The results of these classification methods are discrete with artifacts; however, the overall performances are better than the regression networks. Recently, the classification–regression scheme has emerged with state-of-the-art performances [23], [50], [51]. They propose to use adaptive bins learned from the image to reflect the real distribution of ground truth values of each image and then predict the depth values by a weighted average of the learned bins. Compared to ordinal regression networks, such as DORN [49], classification–regression networks can adapt to different input images and output continuous depth maps.

III. METHODOLOGY

As shown in Fig. 1, the proposed HTC-DC Net consists of three parts, the backbone network to extract features from input images, the HTC-AdaBins module to conduct the HTC as well as incorporate local and holistic information, and the hybrid regression module to get the final height predictions. The proposed network follows the classification–regression paradigm. Based on the extracted features, the HTC-AdaBins module conducts the classification of pixels into bins that are adaptive to each input image, and the hybrid regression module smooths the discrete bins into the continuous output space. The three components are described in detail in this section.

A. Backbone Network

Instead of directly using an encoder–decoder structure for height prediction, the backbone network is used for the generation of feature maps $\{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4, \mathbf{F}_5\}$ from input images $\mathbf{I} \in \mathbb{R}^{3 \times H_0 \times W_0}$, which contain rich spatial and spectral information. Inspired by networks, e.g., U-Net [27], where the intermediate features are aggregated in the later stages of the networks, and following the advanced design in hybrid regression for monocular depth estimation [50], [51], the early injection is done by applying the HTC-AdaBins module and the following hybrid regression process to features of multiple stages in the decoder network, resulting in predictions at different scales.
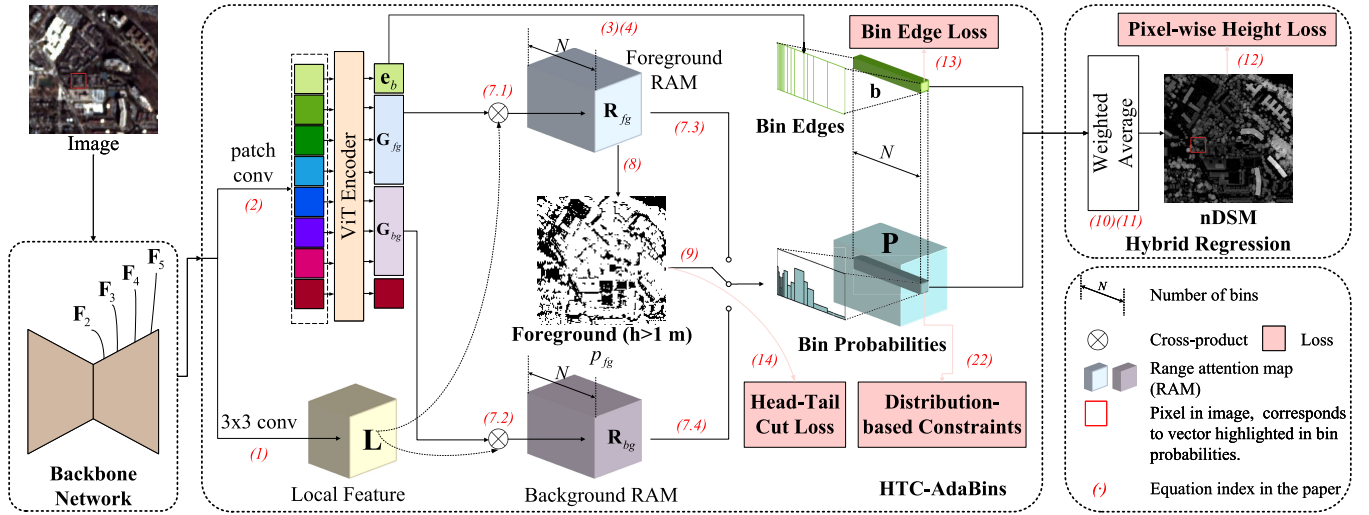
Fig. 1. Network architecture of HTC-DC Net. Following the classification–regression paradigm, the HTC-DC Net is formalized into three parts, the backbone network, the HTC-AdaBins module as the classification phase, and the hybrid regression process as the regression phase. First, a backbone network is used to extract features from images. Based on the features, the HTC-AdaBins module derives the bin edges, which serve as the discretization of the height value range into adaptive bins as classes, and the bin probabilities, regarded as the class probabilities. Finally, the hybrid regression process converts the discretized output space back to a continuous output space by a weighted average of the bin centers according to the bin probabilities. As the contributions of this article, the HTC in the HTC-AdaBins module is used to treat FG and BG pixels separately to account for the long-tail effect in monocular height estimation, and DCs are applied to the predicted bin probabilities for regularization. The FG refers to pixels higher than 1 m. The red numbers in parentheses refer to the corresponding equations.
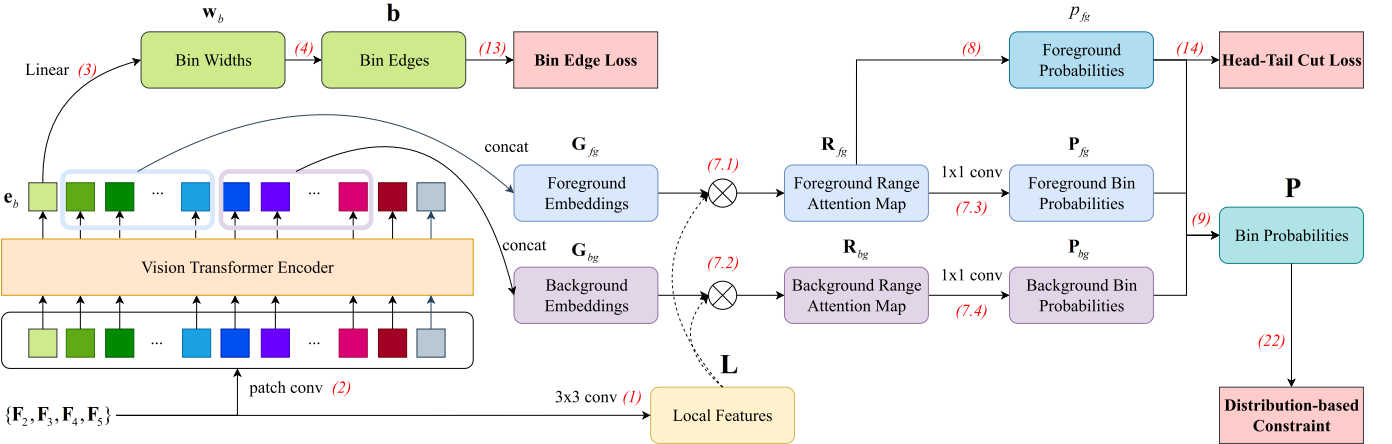

Fig. 2. HTC-AdaBins module contains two branches, namely, the local branch and the global branch. The local branch is responsible for computing local features with a convolutional layer, while the global branch is involved with a ViT encoder for capturing global context. The embeddings from the ViT encoder are utilized for computing the bin edges, the FG bin probabilities, and the BG bin probabilities. During the computation of bin probabilities, a cross-product of the local features and the embeddings from the ViT encoder is conducted to incorporate features of different scopes. The HTC is derived from the FG RAM and used to combine bin probabilities maps for FG and BG pixels. The outputs of the HTC-AdaBins module are supervised by a bin edge loss, an HTC loss, and a DC. The red numbers in parentheses refer to the corresponding equations.

The results of intermediate levels are not taken as the final output, however, used for the computation of loss functions for training.

B. HTC-AdaBins

The HTC-AdaBins module (see Fig. 2) is a variant of the AdaBins module [23] with modifications to address the long-tailed distribution problem in monocular height estimation from remote sensing images. It is used to obtain bin edges $\mathbf{b} \in \mathbb{R}^{N+1}$ and bin probabilities $\mathbf{P} \in \mathbb{R}^{N \times H \times W}$ from the feature maps generated by the backbone network $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$, where N is the number of bins as a hyperparameter, the same for all input images. Intuitively, the bins discretize the continuous

height into classes, which are adaptive to each input image by reflecting the height value distribution of each image, and the bin probabilities serve as the class probabilities, that is, the HTC-AdaBins module converts the regression problem into a classification problem. Besides, the HTC-AdaBins module enables the interaction between local textures learned by the local branch and the global context learned by the global branch. In addition, the HTC enables different treatment of the foreground (FG) and the BG pixels such that the performances for FG and BG pixels are balanced.

1) Local and Global Branch: The local branch with one convolutional layer exploits the local feature pattern $\mathbf{L}$ as

$$\mathbf{L} = \mathrm{conv}_{3 \times 3}(\mathbf{F}) \tag{1}$$

while the global branch with a ViT encoder [22] models the global context. To be fed into the global branch, the feature maps are divided into patches, among which the relations are modeled to refine the embeddings. The process is denoted as

$$\mathbf{E} = \left\{ \mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{\frac{HW}{p^2}} \right\} = \text{ViT}(\text{conv}_{p \times p}(\mathbf{F})) \qquad (2)$$

where p denotes the patch size. The resulting embeddings $\mathbf{E}$ are taken for different uses. To obtain the bins, the first embedding $\mathbf{e}_b := \mathbf{e}_1$ is regarded as the bin width embedding. It is fed into a linear layer fc and then normalized by a softmax function to get the relative bin widths $\mathbf{w}_b \in \mathbb{R}^N$, defined as

$$\mathbf{w}_b = \text{softmax}(\text{fc}(\mathbf{e}_b)). \qquad (3)$$

Finally, given the minimal and the maximal possible values of heights, $h_{\min}$ and $h_{\max}$, the bin edges $\mathbf{b}$ can be obtained by

$$\begin{aligned} \mathbf{b}_0 &= h_{\min} \\ \mathbf{b}_i &= \mathbf{b}_{i-1} + \mathbf{w}_{bi} \quad \forall i = 1, 2, \ldots, N. \end{aligned} \qquad (4)$$

A fixed number m of the embeddings following the first one are concatenated and taken as the global feature $\mathbf{G} :=$ concatenate $(\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_{m+1})$. The global feature $\mathbf{G}$ from the global branch is incorporated with the output from the local branch by a cross-product as follows:

$$\mathbf{R} = \mathbf{L} \times \mathbf{G} \qquad (5)$$

to compute the range attention maps (RAMs) $\mathbf{R}$, which represent the extent of how the height value distribution of a local area compares to the global distribution.

The RAMs $\mathbf{R}$ are then convolved and normalized to get the bin probability maps $\mathbf{P}$ as

$$\mathbf{P} = \text{softmax}(\text{conv}_{1 \times 1}(\mathbf{R})). \qquad (6)$$

2) HTC to Combat the Long-Tailed Effect: As mentioned above, in remote sensing, the height values are usually extremely long-tailed distributed (see Fig. 3), so the majority BG pixels may disturb the computation of RAMs. To mitigate this effect, we propose using an HTC to separate the computation of FG and BG pixels, where FG pixels are defined as pixels with height values greater than 1 m. The definition of the threshold is proved to be reasonable through experiments.

The separation takes effect within the ViT encoder from the global branch. Instead of computing a unique RAM, two different sets of tokens of the same number m, $\mathbf{G}_{\text{fg}} = $ concatenate $(\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_{m+1})$, and $\mathbf{G}_{\text{bg}} = $ concatenate $(\mathbf{e}_{m+2}, \mathbf{e}_{m+2}, \ldots, \mathbf{e}_{2m+1})$, are selected to compute the FG and BG RAMs, $\mathbf{R}_{\text{fg}} \in \mathbb{R}^{N \times H \times W}$ and $\mathbf{R}_{\text{bg}} \in \mathbb{R}^{N \times H \times W}$, and then the bin probabilities $\mathbf{P}_{\text{fg}}$ and $\mathbf{P}_{\text{bg}}$, respectively. Equations (5) and (6) are rewritten as

$$\begin{aligned} \mathbf{R}_{\text{fg}} &= \mathbf{L} \times \mathbf{G}_{\text{fg}} \\ \mathbf{R}_{\text{bg}} &= \mathbf{L} \times \mathbf{G}_{\text{bg}} \\ \mathbf{P}_{\text{fg}} &= \text{softmax}(\text{conv}_{1 \times 1}(\mathbf{R}_{\text{fg}})) \\ \mathbf{P}_{\text{bg}} &= \text{softmax}(\text{conv}_{1 \times 1}(\mathbf{R}_{\text{bg}})). \end{aligned} \qquad (7)$$

In this way, on the one hand, the embeddings from the ViT are utilized in a more efficient way so as to take full advantage of the holistic information acquired by the great computational
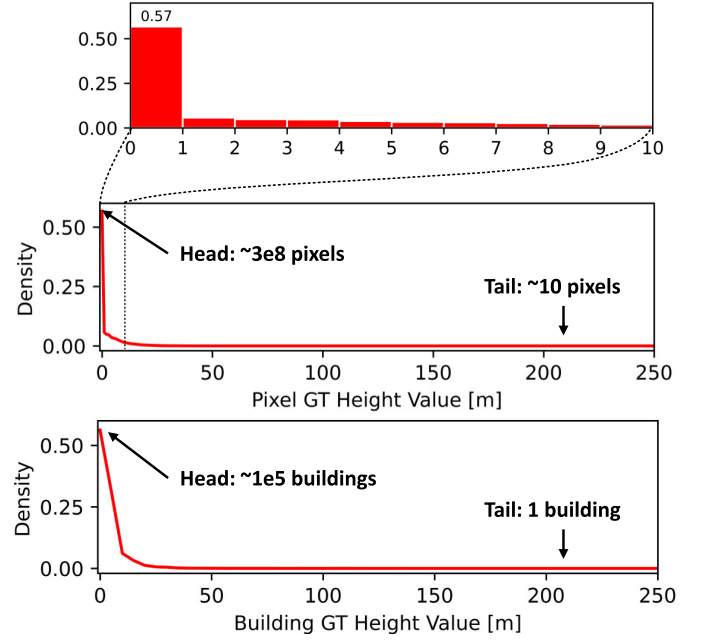

Fig. 3. Height value distribution of GBH training and validation set. The BG with height values smaller than 1 m consists of around 3e8 pixels, which accounts for 57% of the total pixels, while the pixels with very large height values only count to approximately 10 for each 1-m bin. The long-tailed distribution also exists in building height values.

effort, and on the other hand, FG and BG are perceived earlier in the global attention phase, resulting in more distinguishable treatments for them.

The extreme height distribution in the physical world renders the FG and BG pixels each about half of the whole dataset. The HTC problem makes an almost balanced binary classification setting, which is done by simply adding a binary classification head on the FG RAMs. The probability that pixels belong to the FG is computed as

$$p_{\text{fg}} = \text{sigmoid}(\mathbf{R}_{\text{fg}}). \qquad (8)$$

The probability map p_{fg} serves as a mask to combine the bin probabilities computed for FG and BG pixels, which is written as

$$\mathbf{P} = (p_{\text{fg}} > 0.5) \cdot \mathbf{P}_{\text{fg}} + (p_{\text{fg}} \leq 0.5) \cdot \mathbf{P}_{\text{bg}}. \qquad (9)$$

C. Hybrid Regression Process

The hybrid regression process is designed to combine the learned information for each bin by smoothing the discrete output space derived from the HTC-AdaBins module to a continuous output space.

First, a representative value from each bin, i.e., the bin center $\mathbf{c}$, is calculated as the midpoints between two bin edges with

$$\mathbf{c}_i = \frac{\mathbf{b}_{i-1} + \mathbf{b}i}{2} \quad \forall i = 1, 2, \ldots, N. \qquad (10)$$

The final predicted height map $\mathbf{H} \in \mathbb{R}^{1 \times H \times W}$ is formalized as a weighted average of the N bin centers $\mathbf{c}$ according to the

bin probabilities $\mathbf{P}$, i.e.,

$$\mathbf{H} = \sum_{i}^{N} \mathbf{P}_i \mathbf{c}_i. \tag{11}$$

D. Loss Functions

The loss function is composed of four parts.

1) Pixelwise Height Loss: The pixelwise height loss is defined as the $L1$ loss, which is written as

$$\mathcal{L}_h = \frac{1}{|\mathbf{H}|} \sum L_1(\mathbf{H}, \tilde{\mathbf{H}}) \tag{12}$$

where $|\mathbf{H}|$ denotes the total number of pixels and $\tilde{\mathbf{H}}$ denotes the ground truth height map.

2) Bin Edge Loss: To make sure that the bin edges comply with the distribution of ground truth values, the Chamfer loss [52], which computes the bidirectional distances between two point sets, is utilized to supervise the bin edge predictor, i.e.,

$$\mathcal{L}_b = \text{chamfer}(\mathbf{b}, \text{flatten}(\tilde{\mathbf{H}})). \tag{13}$$

Intuitively, the bin edges and the flattened ground truth height maps are seen as two 1-D point sets with height values as the coordinates. For each point in one set, the closest point in the other set is searched, and the distance between the two points is computed and added to the final loss. On the one hand, the distances to the bin edges force them to come from the height values of the input images. On the other hand, the distances to the height values of pixels encourage the bin edges to spread according to the distribution of pixel height values. When the Chamfer loss is small, the distances between the two point sets are small, i.e., the locations of the bin edges comply with the distribution of height values. Ideally, the bin edges lie at the quantiles of the ground truth height values.

3) HTC Loss: As mentioned in Section III-B, an HTC is conducted in the AdaBins module by a binary classification head. The HTC is supervised by a cross-entropy loss, denoted by

$$\mathcal{L}_{\text{htc}} = \text{cross-entropy}(p_{\text{fg}}, \ \tilde{\mathbf{H}} > 1). \tag{14}$$

4) Distribution-Based Constraint: Conventionally, the single regression process yields a single point estimation as the height prediction. In contrast, the hybrid regression process incorporates a distribution-based approach. In the classification phase, a distribution is specified by the bin centers and the bin probabilities. Trivially, taking the most probable value drawn from the distribution accounts for the final height prediction. However, it would lead to discrete output maps solely with values from the bin centers. To overcome this limitation, in the regression phase, a weighted average [see (11)] of bin centers according to the bin probabilities serves as the smoothing of the bins and, thus, enables continuous output space and approaches the mode of the distribution. From a probabilistic perspective, the weighted average is equivalent to computing the expectation value of the underlying height value distribution, which could be far from the distribution mode without any constraints on the distribution. One special case when the expectation and the mode of distribution are close to each other is when the distribution is symmetric and
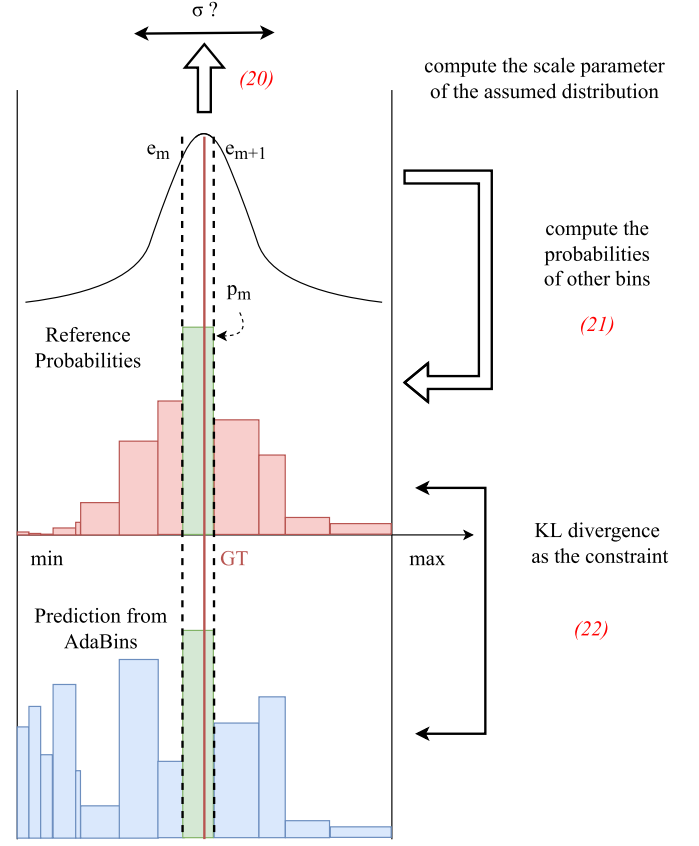

Fig. 4. DC: the constraint is derived and applied in three steps. First, the assumed distribution is solved from the probability of the bin where the GT value lies (the bin bounded by e_m and $e_m + 1$), which is assumed as the mode probability (p_m, in green). With the GT value taken as the mean value, the scale parameter, e.g., the standard deviation of a Gaussian distribution (σ), can be solved analytically. Second, following the assumed distribution, the probabilities for other bins can be calculated as the integral of the PDF within each bin and regarded as the reference bin probabilities (in red). Finally, the reference bin probabilities (in red) are applied as a constraint for the predicted bin probabilities (in blue) in the form of KL divergence. The red numbers in parentheses refer to the corresponding equations.

unimodal, such as a Gaussian distribution. In this case, the hybrid regression process yields the mode value, approximated by the expectation value of the underlying distribution.

Therefore, to enforce that the resulting expectation value of the distribution approaches the distribution mode, a DC is posed on the bin probabilities. As shown in Fig. 4, if the predicted height values obey certain distributions, then the bins are intervals within the defined domain of the distribution, and the bin probabilities are the integrals within the bin intervals. Assuming a known distribution, such as a Gaussian distribution with the ground truth value as the mode, the whole distribution can then be computed and serve as a constraint.

Mathematically, consider that the predicted height h for a pixel is subject to a Gaussian distribution, centered at the ground truth height value $\tilde{h}$, i.e.,

$$h \sim \mathcal{N}(h|\tilde{h}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(h - \tilde{h})^2}{2\sigma^2}\right) \tag{15}$$

where σ as the standard deviation specifies the scale of the distribution, which is unknown and to be solved. The

corresponding cumulative distribution function F is

$$F(h) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{h - \tilde{h}}{\sigma\sqrt{2}}\right)\right) \tag{16}$$

where erf stands for the Gaussian error function, which is written as

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)dt. \tag{17}$$

Then, the probability for the bin where the ground truth value lies, i.e., the mode probability P_m, can be represented as

$$\begin{aligned} P_m &= F(e_{m+1}) - F(e_m) \\ &= \frac{1}{2}\left(\text{erf}\left(\frac{e_{m+1} - \tilde{h}}{\sigma\sqrt{2}}\right) - \text{erf}\left(\frac{e_m - \tilde{h}}{\sigma\sqrt{2}}\right)\right) \end{aligned} \tag{18}$$

where the edges e_m and e_{m+1} make the bin around the ground truth value. When the mode probability is assumed as the reference, (18) has a unique solution of the standard deviation σ, and then, the distribution is fixed. The abovementioned equation can be solved numerically by optimization; however, it leads to large computational burdens. Alternatively, to ease the computation and to solve the equation analytically, it is assumed that the ground truth value lies exactly at the bin center. Then, (18) is simplified as

$$P_m = \text{erf}\left(\frac{e_{m+1} - e_m}{\sigma 2\sqrt{2}}\right). \tag{19}$$

Then, σ can be solved as

$$\sigma = \frac{e_{m+1} - e_m}{2\sqrt{2}\text{ierf}(P_m)} \tag{20}$$

where ierf is the inverse function of the Gaussian error function erf. Even though the assumption is not true, it is necessary to make the problem tractable. Surprisingly, the simplification still brings performance improvements.

After the distribution is fixed by solving the standard deviation σ, the bin probabilities for other bins can be easily computed by

$$P_i = F(e_{i+1}) - F(e_i). \tag{21}$$

Then, these computed probabilities from the assumed distribution are used to supervise the prediction of bin probabilities by the Kullback–Leibler (KL) divergence. The loss can be formulated as

$$\mathcal{L}_{\text{dist}} = \frac{1}{|\mathbf{P}|} \sum -\tilde{\mathbf{P}} \log \frac{\tilde{\mathbf{P}}}{\mathbf{P}} \tag{22}$$

where $\tilde{\mathbf{P}}$ is the probability map from the assumed underlying distribution, with the probability for each pixel in each bin computed by (21).

To consider the differences between FG and BG pixels, different distributions are assumed for them. For BG pixels, the predicted height values h are assumed to follow a uniform distribution, with the ground truth value $\tilde{h}$ as the center point, which is written as:

$$h \sim \text{Uniform}(h|a, b) = \frac{1}{b - a} \tag{23}$$

where a and b are the lower and upper bounds of the distribution, respectively. Similarly, to solve the distribution parameters, the mode probability P_m is set as the reference

$$P_m = F(e_{i+1}) - F(e_i) = \frac{e_{i+1} - e_i}{b - a}. \tag{24}$$

The scale parameter of the distribution denoted as the width $w := b - a$ is derived as

$$w = \frac{e_{i+1} - e_i}{P_m}. \tag{25}$$

Given the distribution is centered at the ground truth height value $\tilde{h}$, then the parameters can be derived as

$$\begin{aligned} a &= \tilde{h} - \frac{e_{i+1} - e_i}{2P_m} \\ b &= \tilde{h} + \frac{e_{i+1} - e_i}{2P_m}. \end{aligned} \tag{26}$$

Finally, the probabilities for other bins can be easily computed using (21).

To compute the total loss, for a single feature map $\mathbf{F}_i$ of level i, the loss function $\mathcal{L}_i$ is defined as the summation of the aforementioned loss function parts, which is denoted by

$$\mathcal{L}_i = \mathcal{L}_h + \mu_1\mathcal{L}_b + \mu_2\mathcal{L}_{\text{htc}} + \mu_3\mathcal{L}_{\text{dist}} \tag{27}$$

where coefficients μ_1–μ_3 are used for balancing between them. To facilitate the multilevel design, the final total loss is the weighted average of total losses for different levels as

$$\mathcal{L} = \sum_{i=1}^{n} \lambda_i \mathcal{L}_i \tag{28}$$

where n is the number of features that are used to compute the loss function and $\{\lambda_i | i = 1, 2, \ldots, n\}$ are coefficients to weight the loss functions of different levels. Normally, the loss functions at later stages should account for more, that is, $\lambda_1 < \lambda_2 < \cdots < \lambda_n$.

IV. EXPERIMENTS

A. Datasets

To demonstrate the efficacy of the proposed network, experiments are conducted on three datasets, Data Fusion Contest 19 (DFC19), Global Building Height (GBH), and ISPRS Vaihingen.

1) Data Fusion Contest 19: DFC19 dataset [53], [54], [55], [56] provides multidate satellite images and ground truth geometric and semantic labels in Jacksonville, Florida, and Omaha, Nebraska, USA. The images cover around 100 km² and date from 2014 to 2016, with a ground sampling distance (GSD) of 1.3 m. The geometric labels are derived from airborne LiDAR data with a nominal pulse spacing of 80 cm. The dataset is delivered as 2783 triplets of images, nDSMs, and semantic maps of size 1024×1024, and GSD of 1.3 m. The semantic maps are processed with only building footprints preserved. To conduct the experiments, the patches are cropped into 44 258 smaller patches of size 256×256 due to GPU memory limits, randomly split into training, validation, and test sets, with 31 152, 4432, and 8944 data samples, respectively.

2) Global Building Height: Existing datasets, including the DFC19 dataset, lack either amount or diversity, so a new dataset—GBH dataset—is proposed and used to demonstrate the efficacy of the proposed network. The GBH dataset is composed of optical remote sensing images from PLANET, height maps in the form of nDSMs, and building footprint maps.

The nDSMs are generated by processing open LiDAR point cloud observations from the authorities. First, the point clouds are denoised, and then, the height values of all points and the height values of ground points are rasterized into DSMs and digital terrain models (DTMs), respectively. Finally, the normalized height is obtained by simply subtracting DTMs from the corresponding DSMs. Besides, building footprint maps are included in the dataset for testing in this article.

The dataset of the current version covers 19 diverse cities around the world and the period from 2013 to 2021. With a patch size of 256×256 and a GSD of 3 m, the dataset is delivered as 20 532 patches, divided into training, validation, and test sets, with 14 971, 3660, and 1901 patches, respectively. Apart from the 19 cities, three cities, Los Angeles, Sao Paulo, and Guangzhou, with 5787 patches, 108 patches, and 1006 patches, respectively, are left out for testing only. It should be noted that only the number of floors for each building is available in Guangzhou, which is converted into buildingwise height by assuming a 3-m floor height.

3) ISPRS Vaihingen: The ISPRS Vaihingen dataset [57], [58] contains aerial orthophotos in IRRG bands, nDSMs generated from LiDAR point clouds, and the corresponding semantic labels, in 33 tiles, with the GSD of 0.09 m. Due to GPU memory limits, they are cropped into patches of size 256×256, randomly split into training, validation, and test sets, with 1209, 279, and 248 data samples, respectively.

B. Evaluation Metrics

To evaluate and compare the performances of different models, the predictions are evaluated in terms of RMSE, RMSE-M, RMSE-NM, and RMSE-B. While RMSE is the pixelwise root-mean-square error for all pixels, RMSE-M measures the RMSE for only building pixels, and RMSE-NM measures the RMSE for only nonbuilding pixels. To further evaluate the capability of the models to generate LoD-1 building models, the buildingwise RMSE, denoted by RMSE-B, is computed with buildingwise predicted values and buildingwise ground truth values, where the buildingwise values are defined as the median of the height values for each building instance represented by one connected component in the building footprint maps.

C. Competitors

The proposed methods are compared to mainstream FCN-based networks, i.e., SegNet [26], FCN [25], U-Net [27], and Efficient U-Net [28]. Methods in this category are mostly designed for semantic segmentation tasks. To facilitate them for height estimation, the final activation layers for classification, e.g., sigmoid or softmax activation, are removed. The output from the last convolutional layer is taken as

the predicted height values. Besides, five networks specifically designed for monocular height estimation are tested for comparison, among which IM2HEIGHT [29], Amirkolaee and Arefi [30], and PLNet [32] are FCN networks taking the problem as a regression task; DORN [48], [49] and Sun et al. [42] and Li et al. [50] covert the regression task into a classification one, with different bin discretization strategies. The architectures of these networks remain unchanged.

D. Implementation Details

The backbone network is built on U-Net [27] and EfficientNet [59] with [60] as the decoder, which gives decoded features of five levels, $\{F_1, F_2, F_3, F_4, F_5\}$, among which $\{F_2, F_3, F_4, F_5\}$ are fed into the rest of the network. For the HTC-AdaBins module, the features are divided into patches of size 4, the number of bins is fixed as 256, and 256 tokens are selected to generate FG and BG embeddings, which complies with the output channel number of the convolutional layer in the local branch. For the hybrid regression process, Gaussian distributions are chosen as the reference distributions for FG pixels, and uniform distributions are assumed for BG pixels. For the loss function, the loss components are weighted with the factors. We set $\mu_1 = 0.01$, $\mu_2 = \mu_3 = 1$, $\lambda_1 = 0$ (F_1 is discarded), $\lambda_2 = 0.125$, $\lambda_3 = 0.25$, $\lambda_4 = 0.5$, and $\lambda_5 = 1$. The values of the abovementioned hyperparameters are proven to be reasonable through experiments.

The proposed networks are trained with the AdamW optimizer, which is a common optimizer used for training ViTs, while the competitors are trained with the Adam optimizer. For all models, the learning rate is set to 1e-4, and early stopping is applied to avoid overfitting. Practically, if the performance of the network fails to improve for ten epochs, the training terminates.

For more implementation details, please refer to the released code.

V. RESULTS

The quantitative results are shown in Tables I–IV. The qualitative results are presented in Figs. 5 and 6 for DFC19 and GBH. Generally, our proposed networks gain better results compared to the existing methods, mostly by large margins.

A. Data Fusion Contest 19

As shown in Table I, our proposed networks consistently achieve the best metrics on the DFC19 dataset. Among the competitors, Amirkolaee and Arefi [30] made the strongest baseline with the best RMSE on all pixels, U-Net [27] performs the best on building pixels, and DORN [49] demonstrates superior results on building instances. However, they are still behind the results from our proposed HTC-DC Nets. For instance, HTC-DC Net B7 outperforms them by margins of 0.7525, 4.2589, and 0.9049 m.

In addition to better quantitative results, our HTC-DC Nets exhibit better preserved minor structures and boundaries and more accurate predictions. As shown in the first qualitative result in Fig. 5, while other networks predict height maps where the canopy textures are highly blurred, the HTC-DC
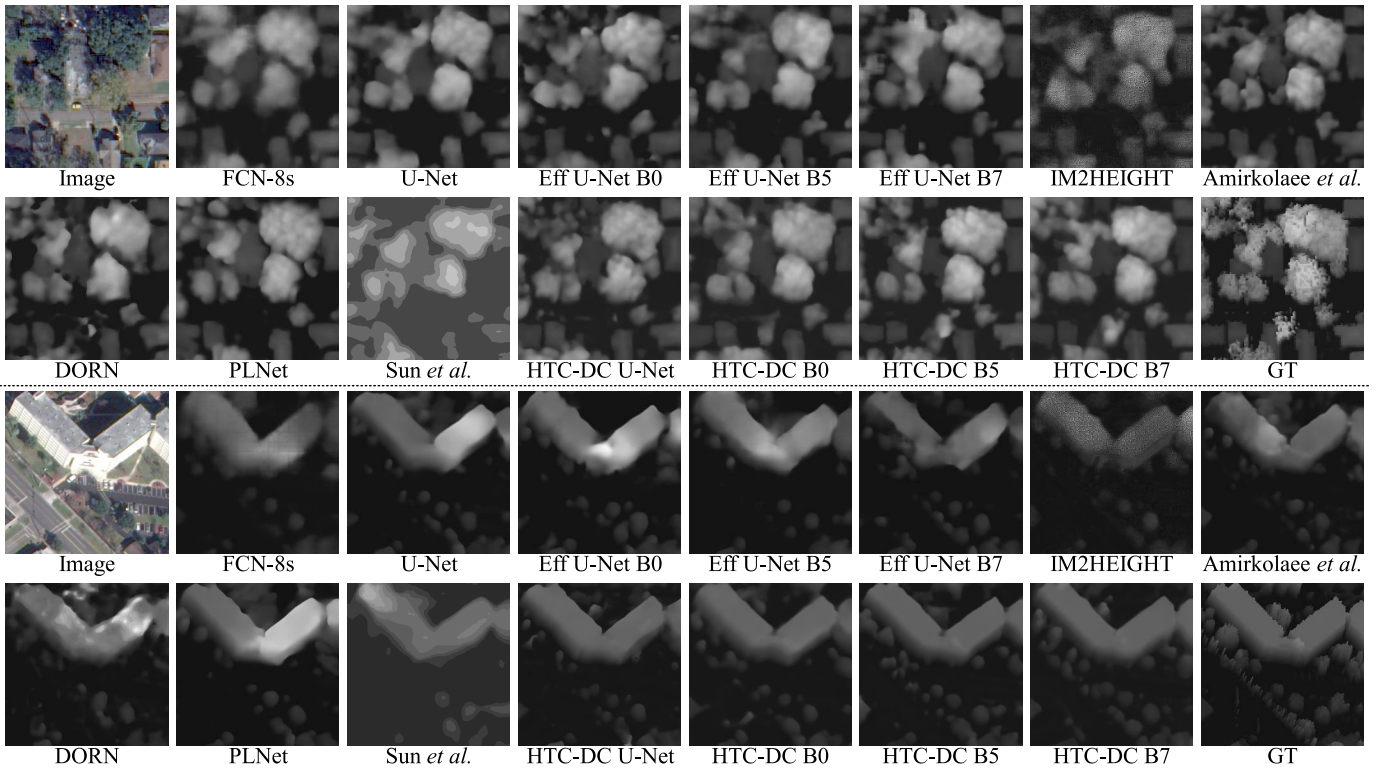
Fig. 5. Qualitative results of different models on DFC19. The maps are scaled to the same range.

TABLE I

EXPERIMENTAL RESULTS OF DFC19. RMSE: PIXELWISE ROOT-MEAN-SQUARE ERROR. RMSE-M: PIXELWISE RMSE FOR BUILDING PIXELS. RMSE-NM: PIXELWISE RMSE FOR NONBUILDING PIXELS. RMSE-B: BUILDINGWISE RMSE. BLOCKS FROM TOP TO BOTTOM: UNIVERSAL FCNS, NETWORKS SPECIFICALLY FOR MHE, AND OUR PROPOSED NETWORKS. COLORS: GREEN, THE BEST METRIC; BLUE, THE SECOND BEST METRIC. HTC-DC U-NET: HTC-DC NET BACKEND BY U-NET [27], HTC-DC NET B{0, 5, 7}, AND HTC-DC NET BACKED BY EFFICIENTNET B{0, 5, 7} [59]

Network	RMSE	RMSE-M	RMSE-NM	RMSE-B
SegNet [26]	7.1397	14.6632	4.9302	5.0540
FCN-32s [25]	4.4580	7.6650	3.6814	3.8626
FCN-16s [25]	3.8230	7.2371	2.9113	3.6513
FCN-8s [25]	3.8070	7.5295	2.7626	3.7290
U-Net [27]	2.9776	5.7762	2.2098	3.3759
Eff U-Net B0 [28]	3.6316	7.6800	2.3965	3.6774
Eff U-Net B5 [28]	3.3220	6.5484	2.4203	3.6403
Eff U-Net B7 [28]	3.3248	6.8710	2.2751	3.5638
IM2HEIGHT [29]	4.8139	10.5586	2.9693	4.1436
Amirkolaee et al. [30]	2.8709	5.7860	2.0345	3.2337
DORN [48], [49]	3.5755	6.4091	**1.9402**	3.2285
PLNet [32]	3.2484	6.6785	2.2397	3.5126
Sun et al. [50], [42]	5.9381	9.3009	5.1894	5.2313
HTC-DC U-Net	3.3886	7.1287	2.2555	3.5954
HTC-DC Net B0	2.4894	4.3801	2.0212	3.2484
HTC-DC Net B5	**2.1813**	**3.8725**	**1.7588**	**2.7850**
HTC-DC Net B7	**2.1184**	**1.5173**	2.1785	**2.3236**

TABLE II

EXPERIMENTAL RESULTS OF GBH CITIES SEEN DURING TRAINING

Network	RMSE	RMSE-M	RMSE-NM	RMSE-B
SegNet [26]	7.5876	13.7589	5.0450	8.1552
FCN-32s [25]	5.6807	8.7639	4.6273	5.1506
FCN-16s [25]	5.2707	8.0504	4.3304	4.7862
FCN-8s [25]	5.0394	7.6875	4.1447	4.3994
U-Net [27]	4.5784	6.7854	3.8532	3.7871
Eff U-Net B0 [28]	5.4468	8.7739	4.2581	4.8907
Eff U-Net B5 [28]	4.9659	7.6708	4.0406	4.3232
Eff U-Net B7 [28]	4.8612	7.3573	4.0244	4.1122
IM2HEIGHT [29]	5.5630	9.1883	4.2319	5.0642
Amirkolaee et al. [30]	4.9693	7.8766	3.9482	4.4514
DORN [48], [49]	5.4389	7.6557	4.4233	4.3266
PLNet [32]	4.9334	7.7507	3.9529	4.2653
Sun et al. [50], [42]	7.6401	10.1139	6.9038	5.9344
HTC-DC U-Net	4.5647	6.5980	3.9119	**3.5237**
HTC-DC Net B0	4.5200	6.6479	3.8257	3.5321
HTC-DC Net B5	**4.4533**	**6.4433**	**3.8138**	**3.4391**
HTC-DC Net B7	**4.4860**	**6.5848**	**3.8025**	3.6158

consistent elevation maps to input images with sharper building boundaries. Besides, the rooftop of the building should be smooth as seen in the predictions from HTC-DC Nets and the ground truth, while other networks predict different heights for the two parts of the "L"-shaped building, and the heights are either overestimated or underestimated.

B. Global Building Height

Generally, larger variabilities are observed on the GBH dataset, which is a more complex and challenging dataset

Nets' predictions are the closest to the ground truth map. In the second example, a tall building is shot with an oblique angle, so the facade is captured in the image. The HTC-DC Nets better distinguish between facade surfaces and roofs, producing

TABLE III
EXPERIMENTAL RESULTS OF GBH LEFT-OUT TEST CITIES

| Network | Los Angeles | | | | Sao Paulo | | | | Guangzhou |
	RMSE	RMSE-M	RMSE-NM	RMSE-B	RMSE	RMSE-M	RMSE-NM	RMSE-B	RMSE-B
SegNet [26]	4.1720	6.4571	**3.2450**	3.8522	12.2260	14.9767	9.8720	13.4439	18.1429
FCN-32s [25]	4.0610	4.2778	3.9975	2.3824	10.6806	13.2142	8.4847	11.6005	15.8830
FCN-16s [25]	3.7616	4.2358	3.6161	2.4009	10.6786	13.3031	8.3832	11.7265	15.8745
FCN-8s [25]	3.6443	4.1717	3.4806	2.2986	10.4170	13.0296	8.1198	11.5194	16.2267
U-Net [27]	3.5821	4.1806	3.3935	2.2906	10.1971	12.4664	8.2600	10.5967	14.3027
Eff U-Net B0 [28]	3.7321	4.5192	3.4771	2.5061	10.6661	13.2319	8.4344	11.5874	16.4835
Eff U-Net B5 [28]	3.6993	4.1933	3.5470	2.3577	10.4470	12.9024	8.3236	11.2397	15.2812
Eff U-Net B7 [28]	3.4688	3.9907	3.3061	2.2500	10.1528	12.4475	8.1869	10.5511	14.3158
IM2HEIGHT [29]	3.7083	4.6525	3.3937	2.5470	10.7472	13.3169	8.5154	11.6647	16.6797
Amirkolaee *et al.* [30]	3.4889	4.1646	**3.2725**	2.3188	10.6489	13.3401	8.2777	11.6855	16.1313
DORN [48], [49]	3.6347	4.1038	3.4840	2.1486	11.6384	13.4164	8.3278	11.8950	15.5949
PLNet [32]	3.6135	4.1601	3.4430	2.2698	10.7689	13.5445	8.3100	11.9087	15.8127
Sun *et al.* [50], [42]	6.2638	4.8582	6.6077	3.8527	10.4421	12.5467	8.6835	11.1389	14.4799
HTC-DC U-Net	**3.3568**	**3.6008**	3.2844	**1.7963**	9.8322	12.3838	7.5673	10.7356	12.6199
HTC-DC Net B0	3.5258	3.7310	3.4655	1.9432	9.6818	11.8199	7.8599	**10.0732**	**11.4203**
HTC-DC Net B5	3.4586	3.7382	3.3752	1.9458	**9.5411**	**11.7565**	**7.6311**	10.2585	11.9607
HTC-DC Net B7	**3.3769**	**3.6131**	3.3070	**1.8286**	**9.3007**	**11.4746**	**7.4234**	**9.8424**	**11.8398**

(see Table II). Despite U-Net performing the best on all metrics among the competitors, it is still inferior to our proposed HTC-DC Nets, such as HTC-DC Net B7, by margins of 0.0924, 0.2007, 0.0507, and 0.1714 m. Note that Sun et al. [42] failed to deliver reasonable outputs on the GBH dataset with collapsed classification outputs, probably due to the higher complexity of the dataset.

As for the test results on unseen cities (see Table III), it is expected that the networks' performances will degrade in cities with significant domain shifts from the training cities. Given that the training cities are mostly located in Europe and North America, the performances of Sao Paulo and Guangzhou are remarkably worse. However, the absolute performance losses of HTC-DC Nets are relatively smaller than other networks.

In the qualitative results presented in Fig. 6, satisfactory outputs are obtained by all the networks on the test set and Los Angeles. However, HTC-DC Nets excel in preserving the minor structures, such as the shapes of complex buildings. Furthermore, while other networks tend to underestimate the heights of tall buildings, the predicted height value for the tallest building in the first visualization example by HTC-DC Net B5 closely aligns with the ground truth. On Sao Paulo and Guangzhou, other networks show severely degraded performances. For example, in the image sample from Guangzhou, FCNs, Eff U-Nets, IM2HEIGHT, Amirkolaee and Arefi, and PLNet generate height maps where the buildings are almost indiscernible, but our proposed HTC-DC Nets still perform well.

C. ISPRS Vaihingen

As shown in Table IV, our proposed networks outperform the existing methods. The best-performing network, HTC-DC Net B7, surpasses the strongest competitor, U-Net, by margins of 0.1981, 0.1874, 0.1933, and 0.1544 m, on RMSE, RMSE-M, RMSE-NM, and RMSE-B, respectively.

In general, superior performance is observed on the ISPRS Vaihingen dataset in comparison to the DFC19 dataset, and

TABLE IV
EXPERIMENTAL RESULTS OF ISPRS VAIHINGEN

Network	RMSE	RMSE-M	RMSE-NM	RMSE-B
SegNet [26]	3.6613	5.5393	2.7982	4.6567
FCN-32s [25]	2.6990	2.2135	2.8379	2.1119
FCN-16s [25]	2.7020	2.9367	2.6219	2.6181
FCN-8s [25]	2.4646	3.0572	2.2402	2.6190
U-Net [27]	1.4916	1.8532	1.3545	1.6963
Eff U-Net B0 [28]	1.9954	2.3565	1.8640	2.2663
Eff U-Net B5 [28]	1.8513	2.1748	1.7341	2.0609
Eff U-Net B7 [28]	1.8167	2.1821	1.6819	1.9990
IM2HEIGHT [29]	2.0775	2.4483	1.9427	1.9390
Amirkolaee *et al.* [30]	1.7782	2.1016	1.6605	2.0308
DORN [48], [49]	1.8182	1.9255	1.5416	1.8509
PLNet [32]	1.8324	2.1544	1.7157	1.9622
Sun *et al.* [50], [42]	2.3735	2.3491	2.3813	2.1973
HTC-DC U-Net	**1.3623**	**1.7410**	1.2151	1.6264
HTC-DC Net B0	1.4286	1.8389	1.2681	1.6547
HTC-DC Net B5	1.3842	1.8229	**1.2092**	**1.6199**
HTC-DC Net B7	**1.3025**	**1.6658**	**1.1612**	**1.5419**

results on the DFC19 dataset, in turn, are expected to be better than those on the GBH dataset. This observation may be attributed to the resolution differences and the resulting complexity changes between the three datasets. From the experiments, it is concluded that our proposed HTC-DC Nets are advantageous on datasets of various GSDs, namely, the GBH dataset of 3-m GSD, the DFC19 dataset of 1.33-m GSD, and the ISPRS Vaihingen dataset of 0.09-m GSD.

VI. DISCUSSION

A. Classification–Regression Paradigm

The proposed HTC-DC Nets employ the classification–regression paradigm to tackle the monocular height estimation task. The HTC-AdaBins module is responsible for predicting classes and their corresponding probabilities based on the input images. Then, the hybrid regression process combines the predicted classes and class probabilities to obtain the final predictions in the continuous output space.
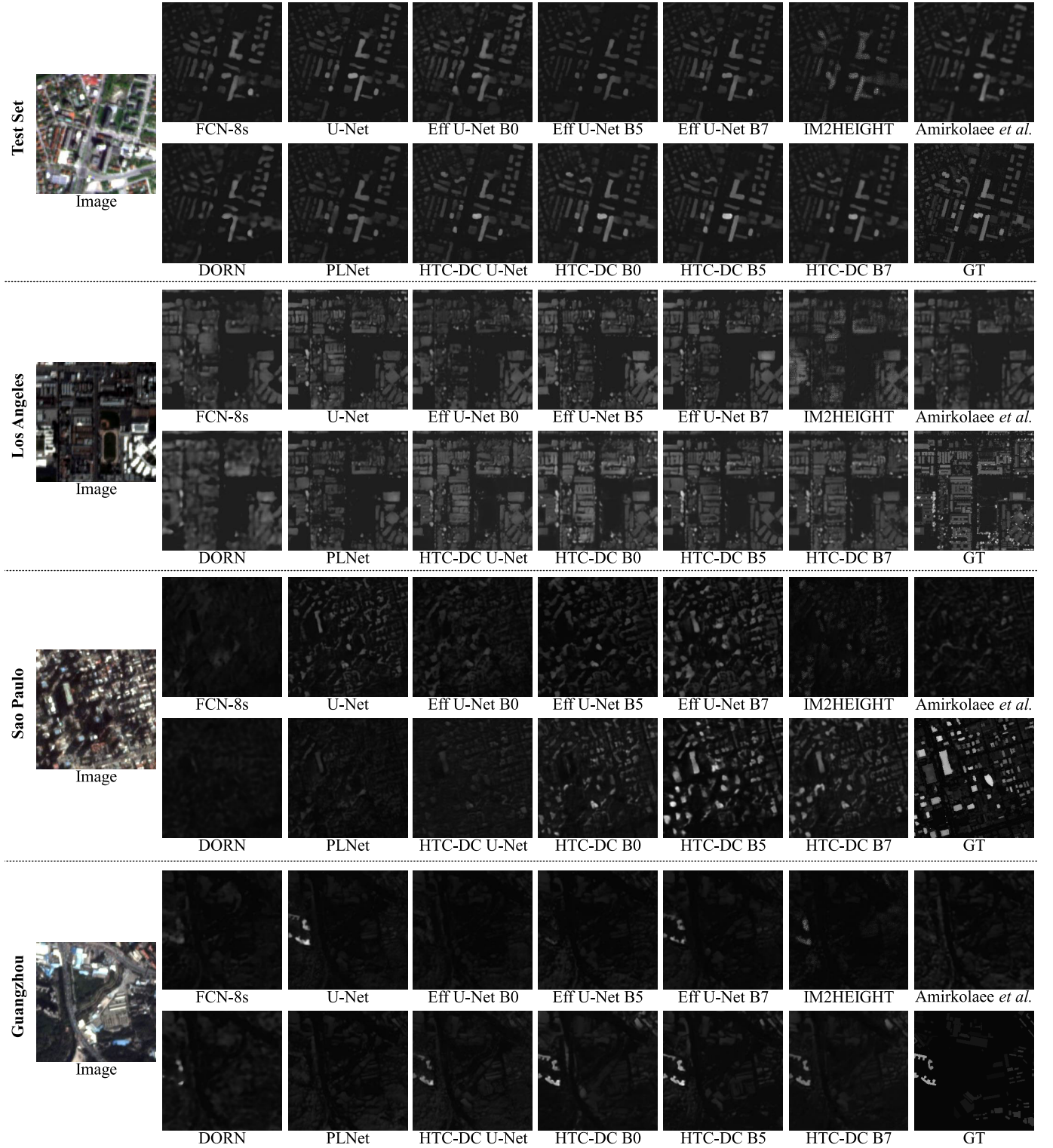
Fig. 6. Qualitative results of different models on GBH. The maps are scaled to the same range. Sun et al. fails and is, thus, not shown here.

The classification phase differs from ordinary classification tasks due to the relationship between classes in terms of quantity and their definition varying across different images, which poses great challenges. Consequently, using a simple classification head yields suboptimal results. In the regression phase, the weighted average serves as the smoothing of the related classes. Taking only the values from the classification results for output leads to discrete and unrealistic output maps, as observed in the results from DORN. The need for continuous output maps justifies the introduction of the hybrid regression process.

Previous works have predominantly followed the regression paradigm, where the height values are directly mapped. Besides, several works employing the classification paradigm convert the regression problem into a classification problem, bringing improvements in performance but often coming with

TABLE V
ABLATION STUDY ON MULTILEVEL EARLY INJECTION

Setting	Test Set				Los Angeles				Guangzhou
	RMSE	RMSE-M	RMSE-NM	RMSE-B	RMSE	RMSE-M	RMSE-NM	RMSE-B	RMSE-B
$\mathbf{F}_5$ only	4.5971	6.9177	**3.8233**	3.8315	**3.4492**	3.9137	**3.3058**	2.0541	12.7593
$\mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_5$	4.6669	6.8616	3.9511	3.7165	3.5190	3.7368	3.4549	**1.9095**	12.2430
$\mathbf{F}_2, \mathbf{F}_4, \mathbf{F}_5$	4.5677	**6.5485**	3.9365	**3.5469**	3.5765	**3.7216**	3.5343	1.9639	**12.1290**
$\mathbf{F}_3, \mathbf{F}_4, \mathbf{F}_5$	**4.5207**	6.7098	**3.8003**	3.6320	**3.4832**	3.7848	**3.3930**	1.9631	12.3482
$\mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4, \mathbf{F}_5$	**4.5200**	**6.6479**	3.8257	**3.5321**	3.5258	**3.7310**	3.4655	**1.9432**	**11.4203**

TABLE VI
ABLATION STUDY ON HTC. HTC: HEAD-TAIL CUT

Setting	Test Set						Los Angeles						Guangzhou	
	Accuracy	RMSE	RMSE-M	RMSE-NM	RMSE-B	RMSE-BG	Accuracy	RMSE	RMSE-M	RMSE-NM	RMSE-B	RMSE-BG	Accuracy	RMSE-B
w/o HTC	-	4.5610	6.9547	3.7526	3.8340	2.8441	-	3.4500	3.8475	3.3289	2.0360	2.6987	-	12.8314
w/ HTC	**0.7929**	**4.5355**	**6.8738**	**3.7504**	**3.7047**	**2.7922**	**0.7339**	**3.3326**	**3.7958**	**3.1893**	**1.9612**	**2.4739**	**0.5975**	**12.4253**

TABLE VII
ABLATION STUDY ON DCS FOR BUILDING PIXELS. # OF IMPROVEMENTS: THE INDICATOR TO SHOW WHETHER A DISTRIBUTION COMBINATION WORKS WELL, DEFINED AS THE NUMBER OF METRICS WHERE IT BRINGS IMPROVEMENTS COMPARED TO THE EXPERIMENT WITHOUT USING ANY DISTRIBUTION CONSTRAINT. THE MAXIMUM # OF IMPROVEMENTS POSSIBLE IN THIS TABLE IS 7

Foreground Distribution	Background Distribution	Test Set		Los Angeles		Sao Paulo		Guangzhou	# of Improvements
		RMSE-M	RMSE-B	RMSE-M	RMSE-B	RMSE-M	RMSE-B	RMSE-B	
None	None	6.8738	3.7047	3.7958	1.9612	12.2859	10.5866	12.4253	-
Uniform	None	6.5964	3.5745	3.7545	2.0248	11.8237	10.1387	11.7504	6
	Uniform	6.6901	3.5919	3.7656	1.9524	11.8537	10.1878	12.4098	**7**
	Gaussian	6.6632	3.5927	3.6640	1.8757	11.8343	10.2058	11.5420	**7**
	Laplace	6.6833	3.5889	3.7223	1.8896	11.9200	10.3905	12.4644	6
	Delta	6.6525	3.5962	3.6465	1.8986	12.0077	10.2301	12.4980	6
Gaussian	None	6.8569	3.6763	3.7828	1.9475	12.1841	10.6231	13.2616	5
	Uniform	6.6479	3.5321	3.7310	1.9432	11.8199	10.0732	11.4203	**7**
	Gaussian	6.6929	3.6799	3.7553	1.9212	**11.7191**	9.9530	11.9169	**7**
	Laplace	6.9970	3.7612	3.6898	**1.8303**	11.9626	10.4068	12.9592	4
	Delta	6.7353	3.6800	3.8820	1.9146	11.6913	10.2209	12.5777	5
Laplace	None	6.4569	**3.4947**	3.6844	2.0072	11.8162	**9.8691**	11.4248	6
	Uniform	7.1464	3.8750	3.7440	1.9064	11.8517	10.2284	12.6783	4
	Gaussian	6.7454	3.5648	3.6051	1.8914	11.7454	10.1551	**11.1319**	**7**
	Laplace	6.5990	3.6187	**3.5820**	1.8972	12.1580	10.5920	12.4754	5
	Delta	6.9813	3.9177	3.7771	1.9472	11.8825	10.1673	13.0317	4
Delta	None	6.7221	3.6314	3.8287	2.2092	11.8509	10.2089	11.5422	5
	Uniform	6.7916	3.7792	3.8050	2.1537	11.9847	10.3700	13.2253	3
	Gaussian	6.6732	3.6277	3.7236	2.0684	11.8418	10.3294	12.1994	6
	Laplace	6.8176	3.7293	3.7665	2.1376	12.1356	10.5050	12.1119	5
	Delta	**6.4508**	3.5147	3.6891	2.0501	11.9219	10.1786	11.8819	6

manually introduced artifacts. In our experiments, we compare our proposed methods to existing works that follow these paradigms. While almost all the previous works follow the regression paradigm, DORN [49] and Sun et al. [42] follow the classification paradigm. Our proposed networks outperform them by significant margins, highlighting the effectiveness of the classification–regression paradigm for monocular height estimation.

B. HTC-AdaBins

Our proposed HTC-DC Nets are built upon U-Net [27] and EfficientNet [59]. By comparing the results of U-Net, Eff U-Net [28], and our proposed HTC-DC Nets, we can demonstrate the efficacy of the HTC-AdaBins module. Notably, our proposed HTC-DC Nets outperform U-Net and Eff U-Nets,

particularly for building pixels. This demonstrates that the adaptation to different input images, as well as the incorporation of local and hostile information, addresses the long-tailed effect, alleviates the underestimation issues, and enhances the performance of monocular height estimation for building pixels.

C. Ablation Studies

Ablation studies are conducted to show the effectiveness of each design component, with the results reported on the GBH dataset.

1) Multilevel Early Injection: Multilevel features are utilized, so height maps of different scales are predicted for supervision, allowing for supervisory signals to occur earlier in the network. In our proposed HTC-DC Nets, EfficientNet
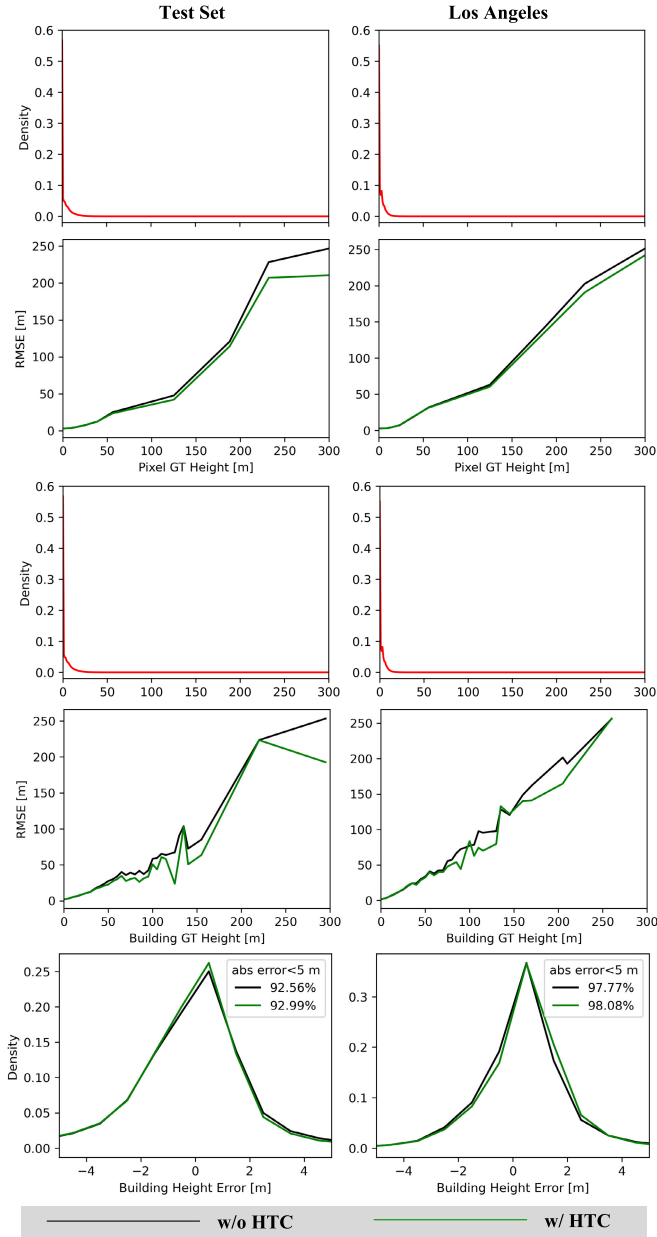
Test Set   **Los Angeles**

Fig. 7. RMSE distribution shows that the HTC helps mitigate the long-tailed effect. From top to bottom: pixel GT height distribution, pixel RMSE versus pixel GT height, building GT height distribution, building RMSE versus building GT height, and building height error distribution.
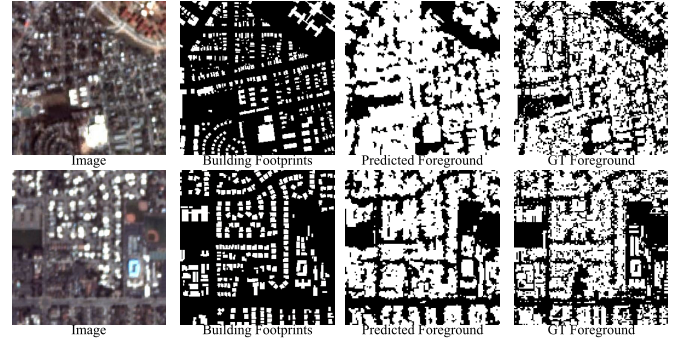

Fig. 8. Results from the HTC and the corresponding building footprint masks. The predicted FG maps are close to the ground truths. For areas with few nonbuilding ground objects (as the second example), the predictions comply with the corresponding building footprint maps.
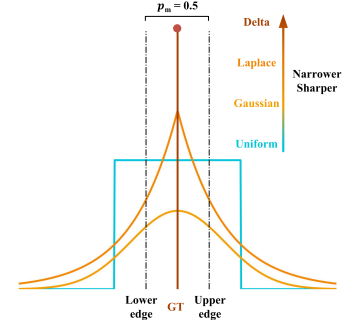

Fig. 9. Distributions considered in the ablation study. The same mode probability is assumed for comparison.

[59] serves as the backbone, and the decoder based on [60] generates features maps at five levels, $\{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4, \mathbf{F}_5\}$. Typically, the low-level features, such as $\mathbf{F}_1$, are too compact for accurate predictions and are discarded. We select features from $\{\mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4, \mathbf{F}_5\}$ and report their results.

The results in Table V show that using features from all four levels yields superior building-related metrics, often ranking among the top two performers. In addition, it leads to comparable RMSE for all pixels. These findings indicate that fusing features from all stages enhance the networks' performance, especially in improving the accuracy of building height predictions.

2) Head-Tail Cut: To mitigate the negative impact of the majority BG pixels on building height predictions, an HTC is employed. Here, we compare the results of networks with and

without the HTC. The purpose of the HTC is to improve models' ability to accurately predict building heights, particularly for tall buildings.

Table VI shows the impact of the HTC on the models' performance. It is obvious that the HTC contributes to the improvement in all metrics. Fig. 7 presents the distribution of RMSE based on pixel ground truth heights and building ground truth heights. It demonstrates that the models' performance improves significantly for both pixels and buildings with higher values as a result of the HTC. This indicates that the HTC is beneficial for areas where the ground truth heights are higher, leading to improved overall performance. As a consequence, the error distribution is "squeezed" toward 0 m, leading to a higher proportion of buildings with absolute errors smaller than 5 m.

Furthermore, regarding the HTC accuracy, since the HTC involves a nearly balanced binary classification task due to the extreme distribution, the classification accuracy is relatively high, as shown in Table VI. In addition, Fig. 8 presents some visualization results from the HTC, where the predictions are close to the ground truth FG maps. In areas with few nonbuilding ground objects, such as vegetation, the predicted FG map accurately corresponds to the corresponding building footprint maps.

3) Distribution Constraints as Supervision: From a probabilistic perspective, the weighted average denoted in (11) is equivalent to computing the expectation value of the underlying height value distribution. Without any constraints, the

TABLE VIII

ABLATION STUDY ON DCS FOR NONBUILDING PIXELS. # OF IMPROVEMENTS: THE INDICATOR TO SHOW WHETHER A DISTRIBUTION COMBINATION WORKS WELL, DEFINED AS THE NUMBER OF METRICS WHERE IT BRINGS IMPROVEMENTS COMPARED TO THE EXPERIMENT WITHOUT USING ANY DISTRIBUTION CONSTRAINT. THE MAXIMUM # OF IMPROVEMENTS POSSIBLE IN THIS TABLE IS 6

Foreground Distribution	Background Distribution	Test Set		Los Angeles		Sao Paulo		# of Improvements
		RMSE-NM	RMSE-BG	RMSE-NM	RMSE-BG	RMSE-NM	RMSE-BG	
None	**None**	3.7504	**2.7922**	**3.1893**	**2.4739**	7.8393	4.7616	-
Uniform	None	3.8793	3.2119	3.5237	3.0009	7.7487	5.6451	1
	Uniform	3.8353	3.1790	3.3873	2.8101	**7.5350**	4.8217	1
	Gaussian	3.8616	3.2373	3.4049	2.9400	7.8226	5.7859	1
	Laplace	3.9288	3.3928	3.5136	3.1256	7.6822	5.0013	1
	Delta	3.9545	3.4338	3.5413	3.1895	7.9198	5.6409	0
Gaussian	None	3.8072	3.0431	3.3036	2.6715	7.7254	**4.4736**	**2**
	Uniform	3.8257	3.1978	3.4655	2.9607	7.8599	5.8695	0
	Gaussian	3.8501	3.1883	3.3374	2.8094	7.8222	6.1055	1
	Laplace	3.8537	3.2602	3.4332	3.0718	7.7656	5.8894	1
	Delta	4.0374	3.5799	3.6969	3.4024	7.6865	5.5319	1
Laplace	None	4.0553	3.5428	3.7021	3.2965	8.0771	6.6759	0
	Uniform	**3.7403**	3.0639	3.4464	3.0200	7.5878	5.1780	**2**
	Gaussian	3.8631	3.3418	3.6327	3.3170	7.7828	5.9134	1
	Laplace	4.1204	3.8706	3.9178	3.8048	7.6909	5.3824	1
	Delta	3.9372	3.4197	3.5051	3.1671	7.8802	5.7204	0
Delta	None	3.9709	3.3611	3.7854	3.3787	7.7560	5.4105	1
	Uniform	4.0210	3.5332	3.7792	3.3955	7.6475	5.1138	1
	Gaussian	4.0375	3.6006	3.7515	3.4089	7.7033	5.5643	1
	Laplace	3.9850	3.5221	3.7192	3.3321	7.8084	5.5763	1
	Delta	4.1970	3.9896	4.0407	3.9340	7.9399	6.1491	0

TABLE IX

ABLATION STUDY ON DCS FOR ALL PIXELS. RMSES FOR ALL PIXELS ARE SHOWN. # OF IMPROVEMENTS: THE INDICATOR TO SHOW WHETHER A DISTRIBUTION COMBINATION WORKS WELL, DEFINED AS THE NUMBER OF METRICS WHERE IT BRINGS IMPROVEMENTS COMPARED TO THE EXPERIMENT WITHOUT USING ANY DISTRIBUTION CONSTRAINT. THE MAXIMUM # OF IMPROVEMENTS POSSIBLE IN THIS TABLE IS 3

Foreground Distribution	Background Distribution	Test Set	Los Angeles	Sao Paulo	# of Improvements
None	None	4.5355	**3.3326**	9.9072	-
Uniform	None	4.5418	3.5759	9.6307	1
	Uniform	4.5386	3.4743	9.5454	1
	Gaussian	4.5488	3.4637	9.6712	1
	Laplace	4.6006	3.5607	9.6479	1
	Delta	4.6096	3.5648	9.8046	1
Gaussian	None	4.5684	3.4151	9.8023	1
	Uniform	**4.5200**	3.5258	9.6818	**2**
	Gaussian	4.5496	3.4340	9.6133	1
	Laplace	4.6408	3.4915	9.7087	1
	Delta	4.6902	3.7385	**9.5345**	1
Laplace	None	4.6254	3.6982	9.7848	1
	Uniform	4.6107	3.5142	9.5690	1
	Gaussian	4.5736	3.6266	9.6076	1
	Laplace	4.7102	3.8462	9.7730	1
	Delta	4.6922	3.5669	9.7229	1
Delta	None	4.6405	3.7950	9.6478	1
	Uniform	4.6947	3.7849	9.6645	1
	Gaussian	4.6728	3.7454	9.6183	1
	Laplace	4.6774	3.7297	9.8165	1
	Delta	4.7243	3.9658	9.7712	1

distribution could be arbitrary, which means that the predicted value can often deviate from the mode of the distribution. However, when assuming symmetric unimodal distributions, e.g., Gaussian distribution or Laplace distribution, the predictions should align precisely with the bins where the ground truth values lie.

The choice of distribution assumption depends on which bins are expected to contribute to the final prediction. If the bins closer to the ground truth bins are supposed to be the primary contributors, then symmetric unimodal distributions are assumed to underlie. In the case where only the ground truth bin is considered for the final prediction, a delta distribution is assumed.

The choice of distribution also determines the extent to which the bins contribute. Within the family of symmetric unimodal distributions, the main difference lies in the sharpness
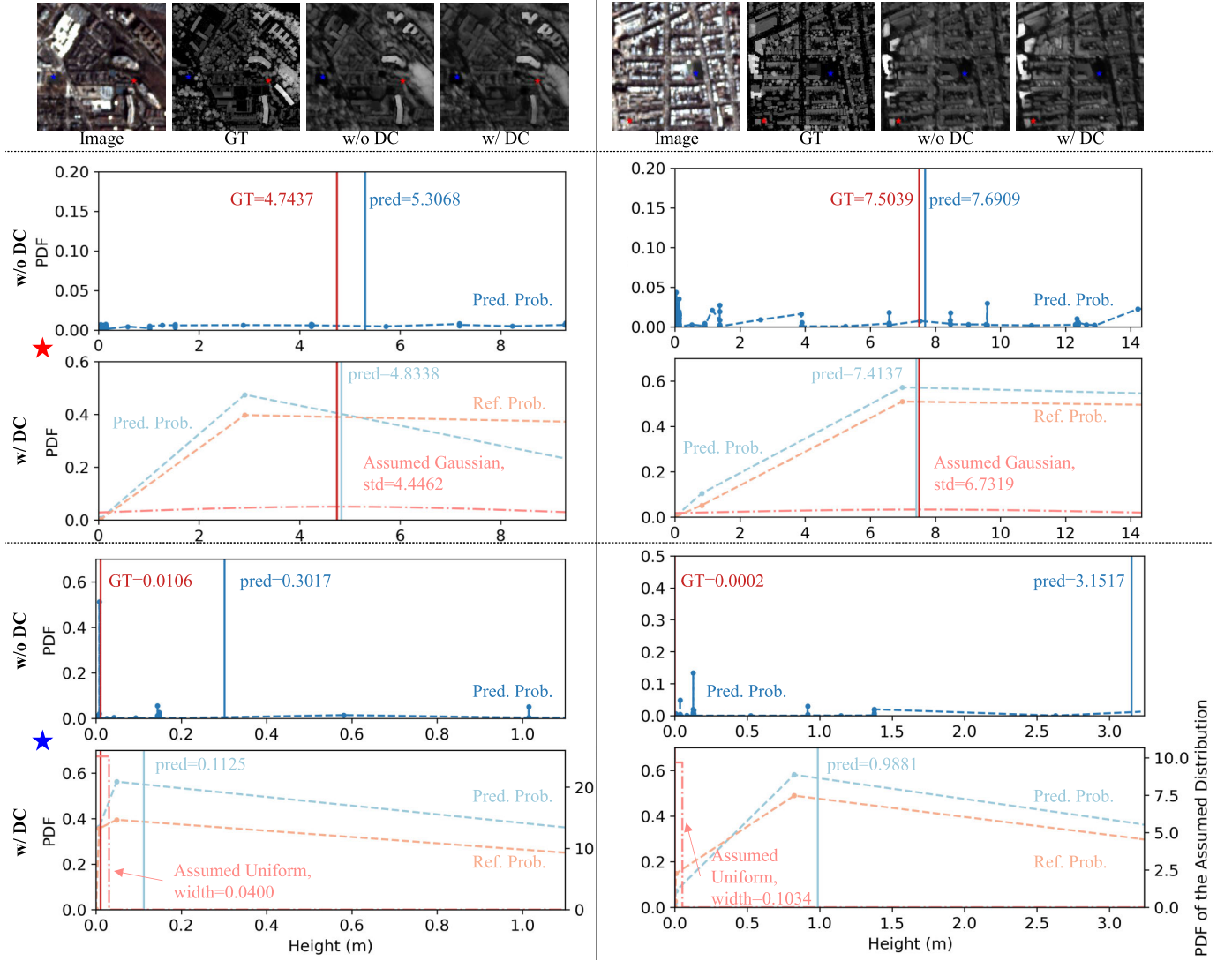
Fig. 10. Visualization of the ablation study for the DCs. Two pixels from the predicted FG and the BG are taken for comparison, which are annotated by the red and blue stars, respectively. With the DCs, the predicted bin probabilities are regularized, so the predicted height values are more accurate.

of the peaks, which represents the margin between the mode probability and the probabilities of the surrounding bins. If all supporting bins are expected to contribute equally, a uniform distribution is assumed.

In the ablation study, four different distributions are implemented to demonstrate the optimality of using Gaussian distributions for the FG and uniform distributions for the BG, in terms of selecting bin contributors and determining their contribution amount. The four distributions used are Gaussian, Laplace, delta, and uniform distributions (refer to Fig. 9). Among these, the delta distribution has the sharpest and narrowest peak, while the uniform distribution has the smoothest and broadest peak. As a complement to Section III-D4, the equation for the derivation of the scale parameter from the mode probability of a Laplace distribution $h \sim \mathcal{L}(h|\tilde{h}, b) = (1/2b)\exp(-(|h - \tilde{h}|/b))$ is provided without proof as

$$b = -\frac{e_{m+1} - e_m}{2\ln(1 - P_m)}. \tag{29}$$

As there is always a compromise between the prediction for building pixels and nonbuilding pixels, the experiments are evaluated threefold. The results are presented in Table VII for building pixels, Table VIII for nonbuilding pixels, and Table IX for all pixels. The indicator of a well-performing distribution is its ability to consistently bring improvements compared to experiments without any distribution constraint, which is represented by the column "# of Improvements" in each table. Based on the results, the combination of Gaussian for FG and uniform for BG yields the largest number of improvements on building pixels (seven) and all pixels (two). Considering that the buildings are of greater interest, this combination is selected as the final configuration. We argue that the choices of distributions have a great effect on the final performances; however, it is hard to analytically decide which distributions to use.

Fig. 10 visualizes the predicted bin probabilities from networks without and with the DC. Visually, the application of the distribution constraints has a subtle effect on the predicted

height maps. However, when examining the bin probability graphs, it is evident that without constraints, the predicted probabilities are relatively small and disorganized, and the predicted values result from a wider range of bin centers. After the constraints are applied, the predicted probabilities are pushed toward the reference bin probabilities derived from the assumed underlying distribution. This indicates that the constraints effectively regularize the bin probabilities and align them with the bins near the ground truth values, as assumed. The bin probability patterns with constraints result in improvements to the hybrid regression results.

It is important to note that there are some failure cases when the mode probabilities are relatively small, leading to extremely large derived standard deviations. This causes the assumed distributions to approach uniform distributions. This phenomenon is likely due to domain shifts, as these failure cases occur more frequently in certain cities.

VII. Conclusion

We propose HTC-DC Net, a network for predicting heights from single remote sensing images. The proposed network utilizes a classification–regression paradigm with a ViT to incorporate holistic features and local features. The regression phase with hybrid regression acts as a smoothing process for the classification phase conducted by the HTC-AdaBins module. With the DCs, the height predictions are efficiently regularized. Besides, to combat the long-tailed distribution problems, a novel HTC is conducted to separate the FGs from the BGs for different treatments. The experiments show that our proposed HTC-DC Net achieves the state-of-the-art performance.

Despite the impressive results given by our proposed HTC-DC Net, the domain shifts between different cities are still challenging for large-scale applications. They lead to performance drops, especially for cities with distinct urban morphologies. Therefore, further works could be done to address the domain shifts by applying domain generalization techniques.

References

[1] H. Arefi and P. Reinartz, "Building reconstruction using DSM and orthorectified images," *Remote Sens.*, vol. 5, no. 4, pp. 1681–1703, Apr. 2013.

[2] T. Partovi, T. Krauß, H. Arefi, M. Omidalizarandi, and P. Reinartz, "Model-driven 3D building reconstruction based on integeration of DSM and spectral information of satellite images," in *Proc. IEEE Geosci. Remote Sens. Symp.*, Jul. 2014, pp. 3168–3171.

[3] Y. Wang, S. Zorzi, and K. Bittner, "Machine-learned 3D building vectorization from satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1072–1081.

[4] A. Matese, S. F. Di Gennaro, and A. Berton, "Assessment of a canopy height model (CHM) in a vineyard using UAV-based multispectral imaging," *Int. J. Remote Sens.*, vol. 38, nos. 8–10, pp. 2150–2160, May 2017.

[5] T. Ota et al., "Aboveground biomass estimation using structure from motion approach with aerial photographs in a seasonal tropical forest," *Forests*, vol. 6, no. 12, pp. 3882–3898, Oct. 2015.

[6] Y. Sadeghi, B. St-Onge, B. Leblon, and M. Simard, "Canopy height model (CHM) derived from a TanDEM-X InSAR DSM and an airborne LiDAR DTM in boreal forest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 1, pp. 381–397, Jan. 2016.

[7] G. Priestnall, J. Jaafar, and A. Duncan, "Extracting urban features from LiDAR digital surface models," *Comput., Environ. Urban Syst.*, vol. 24, no. 2, pp. 65–78, Mar. 2000.

[8] A. F. Elaksher et al., "Reconstructing 3D buildings from LiDAR data," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 34, no. 3, pp. 102–107, 2002.

[9] Z. Chen, H. Ledoux, S. Khademi, and L. Nan, "Reconstructing compact building models from point clouds using deep implicit fields," *ISPRS J. Photogramm. Remote Sens.*, vol. 194, pp. 58–73, Dec. 2022.

[10] X. X. Zhu and R. Bamler, "Superresolving SAR tomography for multidimensional imaging of urban areas: Compressive sensing-based TomoSAR inversion," *IEEE Signal Process. Mag.*, vol. 31, no. 4, pp. 51–58, Jul. 2014.

[11] X. X. Zhu and M. Shahzad, "Facade reconstruction using multiview spaceborne TomoSAR point clouds," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 6, pp. 3541–3552, Jun. 2014.

[12] Y. Sun, Y. Hua, L. Mou, and X. X. Zhu, "CG-Net: Conditional GIS-aware network for individual building segmentation in VHR SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022, Art. no. 5201215, doi: 10.1109/TGRS.2020.3043089.

[13] 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.

[14] Y. Shi, R. Bamler, Y. Wang, and X. X. Zhu, "SAR tomography at the limit: Building height reconstruction using only 3–5 TanDEM-X bistatic interferograms," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 8026–8037, Nov. 2020.

[15] F. Tack, G. Buyuksalih, and R. Goossens, "3D building reconstruction based on given ground plan information and surface models extracted from spaceborne imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 67, pp. 52–64, Jan. 2012.

[16] C. Ginzler and M. Hobi, "Countrywide stereo-image matching for updating digital surface models in the framework of the Swiss national forest inventory," *Remote Sens.*, vol. 7, no. 4, pp. 4343–4370, Apr. 2015.

[17] Z. Xiong, W. Huang, J. Hu, and X. Xiang Zhu, "THE benchmark: Transferable representation learning for monocular height estimation," 2021, *arXiv:2112.14985*.

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

[19] X. Huang and L. Keong Kwoh, "3D building reconstruction and visualization for single high resolution satellite image," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2007, pp. 5009–5012.

[20] M. Izadi and P. Saeedi, "Three-dimensional polygonal building model estimation from single satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 6, pp. 2254–2272, Jun. 2012.

[21] C. Zhao, Q. Sun, C. Zhang, Y. Tang, and F. Qian, "Monocular depth estimation based on deep learning: An overview," *Sci. China Technol. Sci.*, vol. 63, no. 9, pp. 1612–1627, Sep. 2020.

[22] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," Jun. 2021, *arXiv:2010.11929*.

[23] S. F. Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth estimation using adaptive bins," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4008–4017.

[24] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," 2021, *arXiv:2110.04596*.

[25] F. Li et al., "Fully convolutional pyramidal networks for semantic segmentation," *IEEE Access*, vol. 8, pp. 229132–229140, 2020.

[26] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland:: Springer, 2015, pp. 234–241.

[28] B. Baheti, S. Innani, S. Gajre, and S. Talbar, "Eff-UNet: A novel architecture for semantic segmentation in unstructured environment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1473–1481.

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

[30] 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.

[31] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," 2016, *arXiv:1606.00373*.

[32] S. Xing, Q. Dong, and Z. Hu, "Gated feature aggregation for height estimation from single aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[33] Z. Xiong, S. Chen, Y. Shi, and X. X. Zhu, "Disentangled latent transformer for interpretable monocular height estimation," 2022, *arXiv:2201.06357*.

[34] 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.

[35] M. Carvalho, B. Le Saux, P. Trouvé-Peloux, F. Champagnat, and A. Almansa, "Multitask learning of height and semantics from aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1391–1395, Aug. 2020.

[36] M. Elhousni, Z. Zhang, and X. Huang, "Height prediction and refinement from aerial images with semantic and geometric guidance," *IEEE Access*, vol. 9, pp. 145638–145647, 2021.

[37] 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.

[38] 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.

[39] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," Jan. 2018, *arXiv:1703.06870*.

[40] 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)*, Seattle, WA, USA, Jun. 2020, pp. 438–448.

[41] S. Chen, L. Mou, Q. Li, Y. Sun, and X. X. Zhu, "Mask-height R-CNN: An end-to-end network for 3D building reconstruction from monocular remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. IGARSS*, Jul. 2021, pp. 1202–1205.

[42] W. Sun et al., "Rethinking monocular height estimation from a classification task perspective leveraging the vision transformer," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[43] Q. Li et al., "3DCentripetalNet: Building height retrieval from monocular remote sensing imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 120, Jun. 2023, Art. no. 103311.

[44] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," 2014, *arXiv:1406.2283*.

[45] 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)*, Santiago, Chile, Dec. 2015, pp. 2650–2658.

[46] Z. Zhang, Z. Cui, C. Xu, Z. Jie, X. Li, and J. Yang, "Joint task-recursive learning for semantic segmentation and depth estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11214, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 238–255.

[47] K. G. Lore, K. Reddy, M. Giering, and E. A. Bernal, "Generative adversarial networks for depth map estimation from RGB video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake City, UT, USA, Jun. 2018, pp. 1258–12588.

[48] 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.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2002–2011.

[49] 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.

[50] Z. Li, X. Wang, X. Liu, and J. Jiang, "BinsFormer: Revisiting adaptive bins for monocular depth estimation," 2022, *arXiv:2204.00987*.

[51] S. F. Bhat, I. Alhashim, and P. Wonka, "LocalBins: Improving depth estimation by learning local distributions," 2022, *arXiv:2203.15132*.

[52] H. Fan, H. Su, and L. Guibas, "A point set generation network for 3D object reconstruction from a single image," 2016, *arXiv:1612.00603*.

[53] 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)*, Waikoloa Village, HI, USA, Jan. 2019, pp. 1524–1532.

[54] 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)*, Seattle, WA, USA, Jun. 2020, pp. 14500–14508.

[55] 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)*, Nashville, TN, USA, Jun. 2021, pp. 1162–1171.

[56] B. Le Saux, N. Yokoya, R. Haensch, and M. Brown, "2019 IEEE GRSS data fusion contest: Large-scale semantic 3D reconstruction [technical committees]," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 4, pp. 33–36, Dec. 2019.

[57] *2D Semantic Label.—Vaihingen*. Accessed: Aug. 15, 2023. [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx

[58] M. Gerke, "Use of the stair vision library within the ISPRS 2D semantic labeling benchmark (Vaihingen)," Univ. Twente (ITC), Enschede, The Netherlands, Tech. Rep., Dec. 2014, doi: 10.13140/2.1.5015.9683.

[59] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.

[60] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," Mar. 2018, *arXiv:1812.11941*.
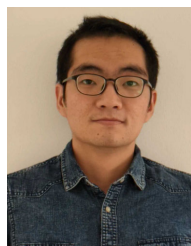
Sining Chen received the bachelor's degree in marine science from Xiamen University, Xiamen, China, in 2018, and the master's degree in Earth-oriented space science and technology (ESPACE) from the Technical University of Munich (TUM), Munich, Germany, in 2020, where he is currently pursuing the Ph.D. degree with the Chair of Data Science in Earth Observation.

He was a DLR/DAAD Doctoral Research Fellow at the Remote Sensing Technology Institute, German Aerospace Center (DLR), 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.

He is currently a Senior Scientist with the Chair of Remote Sensing Technology, TUM. His research interests include fast solver and parallel computing for large-scale problems; high-performance computing and computational intelligence; advanced methods on synthetic aperture radar (SAR) and InSAR processing; machine learning and deep learning for a variety of data sources, such as SAR, optical images, and medical images; and partial differential equation (PDE)-related numerical modeling and computing.

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 with the 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.

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

She is the Chair Professor for Data Science in Earth Observation at Technical University of Munich (TUM) and was the founding Head of the Department "EO Data Science" at the Remote Sensing Technology Institute, German Aerospace Center (DLR). Since May 2020, she has been the Principal Investigator and 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 been the Director of the Munich Data Science Institute (MDSI), TUM. From 2019 to 2022, she was a Co-Coordinator of the Munich Data Science Research School and the Head of the Helmholtz Artificial Intelligence— Research Field "Aeronautics, Space and Transport." She was a Guest Scientist or a Visiting Professor at 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 also a Visiting AI Professor with ESA's Phi-lab, Frascati, Italy. Her main research interests include 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 Sustainable Development Goals (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, the German National Academy of Sciences Leopoldina, and the Bavarian Academy of Sciences and Humanities. She serves on the Scientific Advisory Board in several research organizations, including the German Research Center for Geosciences (GFZ) for the term 2020–2023 and the Potsdam Institute for Climate Impact Research (PIK). She is also an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and *Pattern Recognition*. She serves as an Area Editor for the Special Issue of *IEEE Signal Processing Magazine*.