remote sensing

MDPI

Article

SAR2HEIGHT: Height Estimation from a Single SAR Image in Mountain Areas via Sparse Height and Proxyless Depth-Aware Penalty Neural Architecture Search for Unet

Minglong Xue [1], Jian Li [1], Zheng Zhao [2] and Qingli Luo [1,*]

[1] State Key Laboratory of Precision Measuring Technology and Instruments, Tianjin University, Tianjin 300072, China
[2] Chinese Academy of Surveying and Mapping, No. 28, Lianhuachi Western Road, Haidian District, Beijing 100830, China
* Correspondence: luoqingli@tju.edu.cn

check for updates

Citation: Xue, M.; Li, J.; Zhao, Z.; Luo, Q. SAR2HEIGHT: Height Estimation from a Single SAR Image in Mountain Areas via Sparse Height and Proxyless Depth-Aware Penalty Neural Architecture Search for Unet. *Remote Sens.* **2022**, *14*, 5392. https://doi.org/10.3390/rs14215392

Academic Editors: Riccardo Roncella and Mattia Previtali

Received: 20 September 2022
Accepted: 25 October 2022
Published: 27 October 2022

Abstract: Height estimation from a single Synthetic Aperture Radar (SAR) image has demonstrated a great potential in real-time environmental monitoring and scene understanding. The projection of a single 2D SAR image from multiple 3D height maps is an ill-posed problem in mathematics. Although Unet has been widely used for height estimation from a single image, the ill-posed problem cannot be completely resolved, and it leads to deteriorated performance with limited training data. This paper tackles the problem by Unet with the help of supplementary sparse height information and proxyless neural architecture search (PDPNAS) for Unet. The sparse height, which can be accepted from low-resolution SRTM or LiDAR products, is included as the supplementary information and is helpful to improve the accuracy of the estimated height map, especially in mountain areas with a wide range of elevations. In order to explore the effect of sparsity of sparse height on the estimated height map, a parameterized method is proposed to generate sparse height with a different sparse ratio. In order to further improve the accuracy of the estimated height map from a single SAR imagery, PDPNAS for Unet is proposed. The optimal architecture for Unet can be searched by PDPNAS automatically with the help of a depth-aware penalty term p. The effectiveness of our approach is evaluated by visual and quantitative analysis on three datasets from mountain areas. The root mean squared error (RMSE) is reduced by 90.30% through observing only 0.0109% of height values from a low-resolution SRTM product. Furthermore, the RMSE is reduced by 3.79% via PDPNAS for Unet. The research proposes a reliable method for estimating height and an alternative method for wide-area DEM mapping from a single SAR image, especially for the implementation of real-time DEM estimation in mountain areas.

Keywords: synthetic aperture radar (SAR); height estimation; single SAR image; Unet; sparse height information; proxyless depth-aware penalty neural architecture search (PDPNAS)

1. Introduction

The Digital Elevation Model (DEM) is a 3D representation of the height information about the terrain surface of earth or other planets. It can be mapped from airborne laser radar measurement (LiDAR), photogrammetry, synthetic aperture radar (SAR) stereo measurement, and interferometric synthetic aperture radar (InSAR) technology. DEM generated by LiDAR and InSAR has high accuracy. Limited by the high cost and small coverage, aerial LiDAR products are difficult to be applied in large-scale areas.

Due to its all-time and all-weather ability, SAR is widely applied in earth observations [1]. InSAR is an imperative technology for global Digital Elevation Model (DEM) mapping with the advantage of phase-measuring ability. The classical methods require two or more SAR images to extract elevation. The InSAR images are collected from either two radar sensors imaging at one time or one radar sensor acquired twice with one revisit interval. However, they are limited to the long operation time, high cost, and complex data postprocessing.

By contrast, height estimation from a single SAR image has no restrictions on image acquisition. However, height estimation from a single SAR image is an ill-posed problem, since the pixel value from a certain 2D SAR image can be represented from more than one height value. In order to tackle this problem, geometry-based methods leverage monocular cues, including orientation of the surfaces [2], location motion cues [3] and superpixel segmentation of the planar surfaces [4]. These methods make it possible to construct a 3D structure from a single 2D image. However, complex geometric operations lead to challenges in large memory consumption and long operation time, and it is hard to meet the requirement of real-time performance in practical applications.

With the development of convolution neural networks (CNN), it has been widely used in monocular height estimation from a single aerial image. Typically, a U-shape CNN (Unet) is applied for height estimation from a single imagery. IM2HEIGHT [5] is the first attempt to estimate a height map from a single-view optical image via an end-to-end Unet. Skip connections are proposed to combine the low-level and high-level feature maps for higher height estimation accuracy. Residual blocks are employed to improve the performance of Unet. Costante [6] attempts to use single SAR imagery and a phase map for height estimation. Since the phase map from a single SAR image is uniformly distributed [7] and meaningless, a more efficient postprocessing approach is required to be performed on the phase map for height estimation. Amirkolaee [8] proposes an improved decoder network for Unet. A multi-scale convolution layer is applied in the decoder subnetwork for capturing more context information. The spatial resolution of the output feature map has been improved by this method. IMG2DSM [9] leverages the generative adversarial network for monocular height estimation on the base of Unet. Son [10] proposes a deep monocular depth network for single aerial imagery height estimation. It is especially efficient for 3D reconstruction in urban areas when the building suffers from sudden change. However, the ill-posed one-to-many problem is not completely resolved by Unet, especially with limited data in remote sensing. The accuracy of the estimated height map is not high enough, especially for mountain areas with a wide range of elevations.

Supplementary information enforcing geometric constraints has proved to be efficient for tacking the ill-posed problem. These methods can be categorized into: multi-task learning and multi-sensor fusion methods. In multi-task learning methods, additional visual tasks are jointly trained with height estimation. The additional tasks are composed of pixel-wise semantic segmentation [11,12], 2D/3D edges detection [13] and signed distance prediction [14]. The potential relationships between these tasks are learned by Unet, which leads to superior performance. Mallya [15] proposed the PackNet, and it includes multiple visual tasks with the use of one single Unet. In multi-sensor methods, multi-source data are utilized as the inputs of Unet to estimate the height map. IM2ELEVATION [16] takes both of the LiDAR and optical data to be the inputs of Unet for height estimation and a registration strategy based on mutual information. Both LiDAR and optical data are accepted as the inputs of Unet for height estimation. Amirkolaee [17] suggests that features are not well defined for the non-ground object in a single aerial image. Then, the estimated height map from ground objects and a Shuttle Radar Topography Mission (SRTM) DEM data from non-ground objects are combined for high-precision height estimation. Kim [18] proposes a very deep super resolution (VDSR) for depth completion based on the VGG16 network. Xia [19] leverages the sparse height information for monocular depth estimation (PrDepth). DORN features are extracted from a network which is pretrained on a large-scale depth estimation KITTI dataset. Since there are no available large-scale SAR datasets for extracting suitable DORN features in advance, the DORN features pretrained by depth estimation datasets may not be helpful for height estimation from a single SAR image.

Another approach to tackle this problem by Unet and similar architectures with limited data is to benefit from transfer learning and data augmentation techniques, and they can be used as alternative methods to the supplementary information-based methods.

UNet-VGG16 [20] is proposed to improve the performance of segmentation for magnetic resonance imaging (MRI) images with the help of a pretrained model. The encoder of UNet-VGG16 is extracted from VGG16, which is pretrained on ImageNet datasets, and the correct classification ratio (CCR) has been improved significantly by UNet-VGG16. Pellegrin [21] leverages the depth estimation Unet model pretrained on a large-scale urban dataset KITTI, and then, the pretrained model was applied for a single aerial image height estimation. Few-shot learning, which relies on knowledge transferring, focuses on improving the performance of CNN with limited data. Stan [22] proposes an efficient approach for unsupervised few-shot continual learning. Since the distribution of the source domain is unacceptable in this case, a surrogate Gaussian prototypical distribution estimator is used to measure distances between the data from the source and target domain. Then, the distribution of the two domains can be aligned indirectly. Wibowo [23] proposes using a dynamic adaptive subspace classifier [24] to improve the performance of few-shot learning. Zhang proposes using the data augmentation method to generate more training data for vehicle detection from two-pass SAR images [25]. Since the registration errors exist in the two-pass SAR images and they have a great impact on the detection precision, simulated registration errors are introduced to generate the training data which is more realistic.

Neural architecture search (NAS) has been widely used in automatically designing the optimal architecture for CNN. An over-parameterized network consisting of many mixed blocks is employed to search for the optimal architectures. Each mixed block contains many conventional operations of CNN, including 3 × 3, 5 × 5, 7 × 7 convolution/transposed convolution/down or upsampling, identity mapping, etc. Limited by the computation consumption, conventional NAS methods [26,27] are trained on proxy tasks when they are applied to large-scale datasets. The proxy tasks mean smaller datasets/models, fewer training epochs, etc. Typically, the performance of optimal architecture learned on proxy tasks is not guaranteed on the target task. The emergence of proxylessNAS [28] has made it possible to learn the optimal architecture on a target task directly. Most of the feasible paths linking candidate operation sets are cut off, and only one of them is set as active in the training stage. Therefore, the performance of the optimal model searched by proxyless NAS is guaranteed on the target task, and it can be deployed in practical applications. The searching cost is reduced by path binarization and a path-level pruning in the training stage. The time complexity of proxylessNAS is nearly the same as the conventional networks, and the classification accuracy increases by 2.78% on ImageNet datasets. NAS-Unet [29] is the first attempt to combining the advantages of proxylessNAS and Unet. The optimal architecture learned by NAS-Unet may be asymmetric. It leads to a deteriorated performance for Unet. Another problem of NAS-Unet is that the Unet searched by proxyless may be sub-optimal. Since deeper layers prefer larger kernels according to the results of proxylessNAS on the ImageNet classification task, the relative depth of it is changed due to the existence of skip connections (or cweights in NAS-Unet) in Unet, and it may be paused to select the optimal architecture for Unet.

Inspired by the success of height estimation with multi-source data and NAS-Unet [29], we propose sparse height information as the supplementary input of Unet and proxyless depth-aware penalty neural architecture search (PDPNAS) for Unet to achieve better height estimation accuracy. The sparse height information can be produced from LiDAR sensors [30], Time-of-Flight (ToF) sensors [31], stereo matching [32], UAV photogrammetry [33], etc. Since the spatial resolutions of the height data obtained from the above sensors are different, a parameterization method for generating a sparse height information with the different sparse ratios is proposed. A nearest-neighbor filling method is applied for the sparse height information. The performance of Unet can be improved by observing only a small fraction of height values from other products. In the end, the optimal architecture for Unet is learned by PDNAS without increasing large computation cost. This research will provide an effective solution for a single SAR elevation retrieval and a new idea for real-time DEM mapping.

The key contributions of this work are as follows:

- A height estimation network based on Unet was proposed. Sparse height information **SH** and distance map **d** are used as additional input for higher reconstruction accuracy. The root means square error of height estimation in a mountain area can be improved from ∼315 m to about 32 m (the sparse ratio of **SH** is 0.011%).
- A customized method for generating **SH** with different sparse ratios is proposed. To accommodate for various sparse inputs, a mask function is proposed to simulate the sparse patterns.
- A proxyless depth-aware penalty neural architecture search is proposed to learn the optimal architecture for Unet.

2. Materials and Methods

2.1. The Proposed Method

Since the phase map of a single SAR image is noisy and meaningless, only the intensity of the SAR image will be included as the input of our network. The intensity map of SAR is denoted as $\mathbf{I} \in \mathscr{I}^{M \times N}$, where M and N are the numbers of pixels in azimuth and range directions. The original ground truth DEM is in map coordinates, and a geocoding process is required to convert to radar coordinates. The geocoded ground truth height map in SAR coordinate systems is denoted as $\mathbf{H} \in \mathscr{H}^{M \times N}$. The estimated height map from a single SAR image I is denoted as $\hat{\mathbf{H}} \in \mathscr{H}^{M \times N}$.

Height estimation from a single SAR image is to establish the mapping function $f_\Theta : \mathscr{I} \rightarrow \mathscr{H}$ from the intensity map of SAR imagery to the height domain, where Θ are the parameters of f. The original sparse height is $\mathbf{Hs} \in \mathscr{H}^{M \times N}$. To be comparable with different sparsity, a parameterized method that can generate $\mathbf{H}_S$ with different downsampling factors $S \times S$ is proposed in this paper. $\mathbf{H}_S$ requires being densified by nearest neighbor filling for Unet training. The densified sparse height information is denoted as **SH**. The distance map **d**, which is calculated from nearest neighbor filling densification, is also recorded and accepted as inputs. The loss function $\mathscr{L}$ is listed as following:

$$\mathscr{L} = \min_{\Theta} \sum_{x=1}^{M} \sum_{y=1}^{N} ||\mathbf{H}(x,y) - \hat{\mathbf{H}}(x,y)||_2^2, \tag{1}$$

$$\hat{\mathbf{H}}(x,y) = f_\Theta(\mathbf{I}(x,y), \mathbf{SH}(x,y), \mathbf{d}(x,y)), \tag{2}$$

where (x,y) is the location in azimuth and range coordinates.

All of the inputs and outputs are normalized from 0 to 1 due to the leakyReLu [34] and sigmoid activation function employed in our model. The normalization procedure is listed as the following:

$$\mathbf{I}^{'} = \frac{I}{max(\mathbf{I})}, \tag{3}$$

$$\mathbf{d}^{'} = \frac{\mathbf{d}}{max(\mathbf{d})}, \tag{4}$$

$$\mathbf{SH}^{'} = \frac{\mathbf{SH}}{1.1 \cdot max(\mathbf{SH})}, \tag{5}$$

$$\mathbf{H}^{'} = \frac{\mathbf{H}}{1.1 \cdot max(\mathbf{SH})}, \tag{6}$$

According to Equations (3)–(6), the normalization of $\mathbf{I}^{'}$ is dividing by $\max(\mathbf{I})$, and the normalization of $\mathbf{d}^{'}$ is the same as that of $\mathbf{I}^{'}$. It is slightly different for $\mathbf{H}^{'}$ and $\mathbf{SH}^{'}$; for that, the data range of **H** cannot be obtained in a practical application. Therefore, the normalization of $\mathbf{SH}^{'}$ and $\mathbf{H}^{'}$ is dividing by $1.1 \times max(\mathbf{SH})$ to obtain a valid normalized vector, which is from 0 to 1.

All of the train and test data of Unet are cut into 256 × 256 small slices. Two examples of **I** and its corresponding **H** which measure 256 × 256 pixels are shown in Figure 1. Typically, since the 256 × 256 patches are predicted independently by Unet, post-processing on test sets is applied before connecting the adjacent height patches in other research. For instance, Amirkolaee [8] leverages height shifting to avoid the large distinction among adjacent predicted height patches. In our methods, the supplementary **SH** restricts the estimated range of adjacent patches. Therefore, no additional post-processing is needed to achieve high precision when connecting the small patches to a large complete height map.

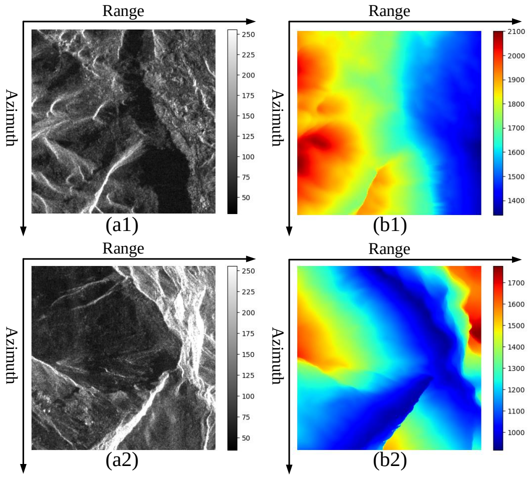


Figure 1. Two examples of acquired SAR (**a1**,**a2**) and its corresponding geocoded ground truth DEM (**b1**,**b2**). The color bar represents the height information, and the unit is meters.

In the following, we will introduce the main steps including coordinate transformation and registration, sparse height information extraction and PDPNAS for Unet.

2.2. Coordinate Transformation and Image Registration

Due to the side-looking imaging geometry of SAR, SAR images and ground truth DEM are located in different coordinates. In order to make it trainable for Unet, ground truth DEM is transformed from map coordinates to SAR coordinates, which is known as geocoding. The straightforward geocoding suffers from inevitable holes in layover/shadow areas, and it occurs more in mountain areas. In this paper, a backward geocoding is employed based on the lookup tables between SAR and DEM coordinates. Details about it can be referred to [35]. The geocoding procedure can be divided into three steps:

1. Establishing initial geometric transformation between SAR image and DEM based on range-Doppler (RD) model [36].
2. Refining the geometric transformation by offset calculation between SAR data and simulated SAR based on DEM.
3. Resampling image data sets from DEM to SAR coordinates system.

The initial definition for geometric transformation between SAR and DEM is based on orbital information and DEM parameters. The initial definition of transformation is not accurate due to the errors of orbital state vectors. Meanwhile, the resolution of DEM and SAR images are different, and it requires interpolation operation. Thus, a refinement procedure is required to achieve high-precision geocoding products.

In the refinement procedure, the registration offsets between the corresponding locations of the geocoded SAR and DEM image are computed. A conventional method is to leverage the manually selected control points to generate many pairs of pixel positions. A more efficient way is to leverage the lookup tables which represents the geometric transformation between two coordinate systems. The lookup table is based on the orbital data and the parameters of SAR image and DEM. A simulated SAR intensity map is generated by the DEM and lookup table. The registration offsets between them are computed automatically through cross-correlation analysis.

Once a high-precision lookup table is established, the geometric transformation can be operated from DEM (map) to an SAR (radar) coordinated system. Due to the existence of layover and shadow in SAR images, an interpolate operation is performed in the areas of layover and shadow. In order to explore the representation ability of our network, comparison experiments are performed in the areas of layover and shadow and beyond.

2.3. Parameterization of Sparse Height

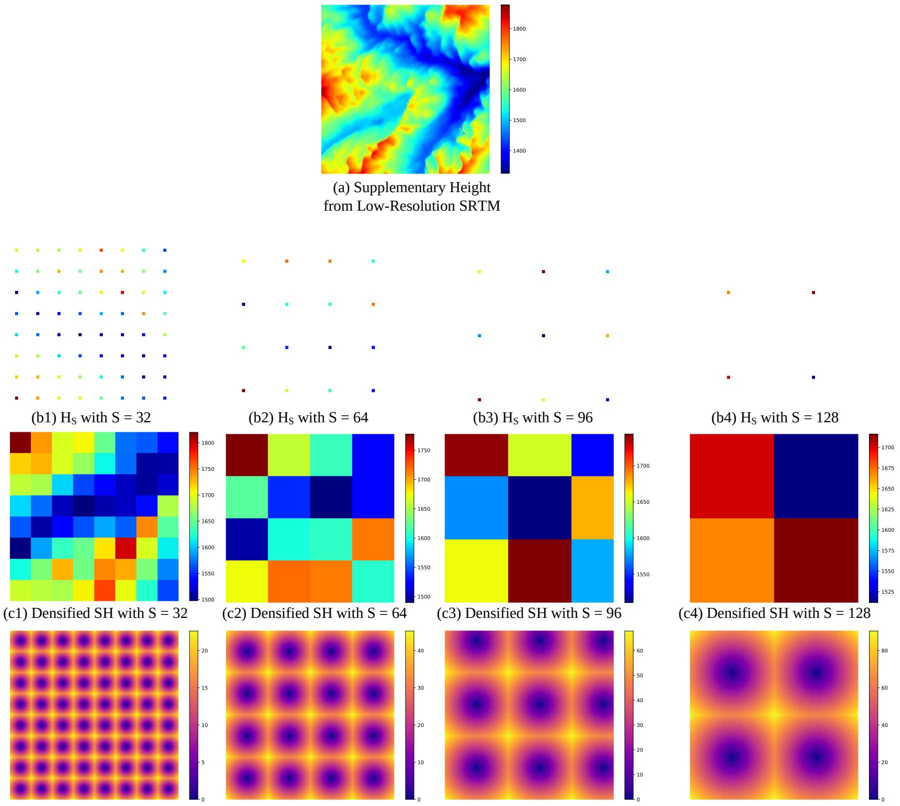
The original sparse height map $\mathbf{H}_S$ can be observed by in-suit measurements (conventional leveling or GNSS measurement), LiDAR [37,38] or downsampled low-resolution SRTM products. In practice, the paired high-resolution 3D LiDAR points and SAR products are hard to acquire. The available LiDAR maps from OpenTopography products measure 80 × 50 pixels in a SAR intensity map measuring 8130 × 5796 pixels. Since the image slices in the train datasets measure 256 × 256 pixels, the limited data obtained from LiDAR products make it impossible for Unet training. In this situation, $\mathbf{H}_S$ is acquired by downsampling from low-resolution SRTM products in this paper.

In order to explore the effect of $\mathbf{H}_S$ with different sparse ratios on single SAR height estimation, we propose to extract the sparse height points by downsampling from a low-resolution SRTM product via a mask mapping function $M_{S\times S}$, where $S \times S$ are the downsampling factors. In order to generate $\mathbf{H}_S$ with different sparse ratios, the whole DEM image is divided into thousands of areas. Each area measures $S \times S$ pixels. Only the middle point of each area is kept, and the others are set as 0. In order to feed $\mathbf{H}_S$ to Unet, a nearest neighbor filling process is required to generate densified sparse height **SH**. Then, the whole **SH** and **d** are cut into 256 × 256 small slices. Several examples of the 256 × 256 small slices are shown in Figure 2a is the supplementary height map which is received from low-resolution SRTM products. Figure 2(b1–b4) are the original sparse height map $\mathbf{H}_S$ with different downsampling factors. Figure 2(c1–c4) are the densified sparse height **SH** generated from Figure 2(b1–b4) by nearest neighbor filling. Figure 2(d1–d4) are the distance maps which represent the Euclidean distance from a pixel without height values ($M_{S\times S}(x,y) = 0$) to its nearest sampling center ($M_{S\times S}(x^*,y^*) = 1$), $d = \sqrt{(x-x^*)^2 + (y-y^*)^2}$. The mathematical expression of $M_{S\times S}$ is as follows:

$$M_{S\times S}(x,y) = \begin{cases} 1, & x = \dfrac{S}{2} + S\cdot i \;\; and \;\; y = \dfrac{S}{2} + S\cdot j, \;\; where \;\; i,j = 0,1,2,\ldots \\ 0, & others. \end{cases} \tag{7}$$

where $(i,j), i,j = 0,1,2,\ldots$ denote the ordinal numbers of the downsampling center points.

Details about the algorithm are shown in Algorithm 1. At first, the mask function $M_{S\times S}$ is established as described in Equation (7). Then, $\mathbf{H}_S$ is generated by the Hadamard product (Hadamard product: For $\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}$, $\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix}$, $\mathbf{X} \circ \mathbf{Y} = \begin{bmatrix} x_{11}\cdot y_{11} & x_{12}\cdot y_{12} \\ x_{21}\cdot y_{21} & x_{22}\cdot y_{22} \end{bmatrix}$) of ground truth height map **H** and mask $\mathbf{M}_{S\times S}$, $\mathbf{H}_S = \mathbf{H} \circ \mathbf{M}_{S\times S}$. It allows obtaining sparse height points with various sparse ratios by setting the downsampling factors $S \times S$. The downsampling factors $S \times S$ are set as 32 × 32, 64 × 64, . . . , and 192 × 192. Taking $S = 96$ as an example, it represents that only one real height point is kept from every 96 × 96 pixels. In order to feed $\mathbf{H}_S$ to Unet, it requires being densified by nearest neighbors filling. The densified sparse height is labeled as **SH**. Euclidean distance maps measuring the distance from $\mathbf{M}_{S\times S}(x,y)$ to the nearest sampling center $\mathbf{M}_{S\times S}(x^*,y^*)$ are stored as $\mathbf{d} = \sqrt{(x-x^*)^2 + (y-y^*)^2}$. Then, the supplementary inputs of Unet are generated by concatenating SAR image **I** with **SH** and **d**.



(a) Supplementary Height from Low-Resolution SRTM

(b1) H_S with S = 32 (b2) H_S with S = 64 (b3) H_S with S = 96 (b4) H_S with S = 128

(c1) Densified SH with S = 32 (c2) Densified SH with S = 64 (c3) Densified SH with S = 96 (c4) Densified SH with S = 128

(d1) distance d with S = 32 (d2) distance d with S = 64 (d3) distance d with S = 96 (d4) distance d with S = 128

Figure 2. Generate original sparse height $\mathbf{H}_S$ with different downsampling factors $S \times S$ from supplementary height map. Each image measures 256 × 256 pixels. To feed it to Unet, $\mathbf{H}_S$ is densified by nearest neighbor filling (NNF). The densified sparse height is **SH** and the distance map from NNF is **d**. S is the downsampling factor. Details about the process is listed in Algorithm 1.

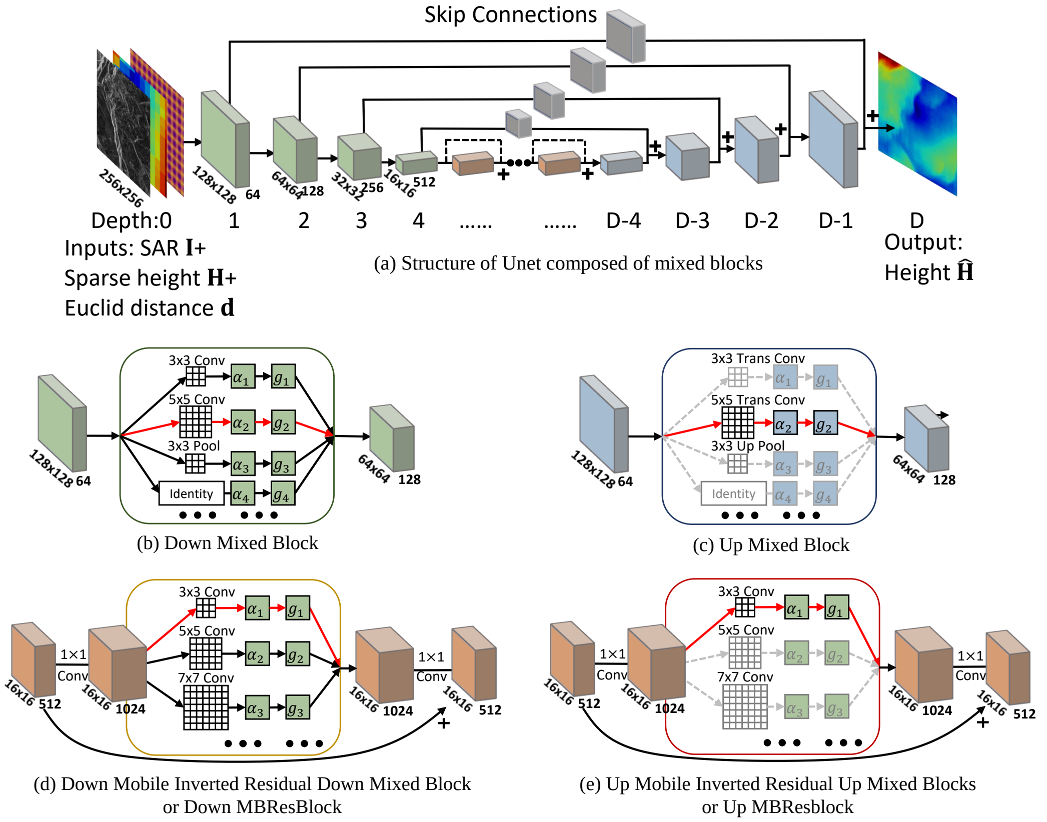
Algorithm 1 Parameterized methods for generating inputs of SAR2HEIGHT.

```
Input: SAR image I, ground truth height map H, downsampling factors $S \times S$
Output: Sparse height information $\mathbf{H}_S$, densified sparse height SH, Euclidean distance map d.
$M \leftarrow$ get Image Height of I
$N \leftarrow$ get Image Width of I
Generate $\mathbf{M}_{S\times S}$ according to Equation (7)
$\mathbf{H}_S = \mathbf{H} \circ \mathbf{M}_{S\times S}$
for $x = 1$ to $M$ do
    for $y = 1$ to $N$ do
        $(i^*, j^*) = \arg\min_{i,j} \sqrt{(x - (\frac{S}{2} + S \cdot i))^2 + (y - (\frac{S}{2} + S \cdot j))^2}$
        $x^* = \frac{S}{2} + S \cdot i^*$        ▷ Get the coordinates of nearest downsampling center.
        $y^* = \frac{S}{2} + S \cdot j^*$
        $\mathbf{SH}(x, y) = \mathbf{H}_S(x^*, y^*)$
        $\mathbf{d}(x, y) = \sqrt{(x - x^*)^2 + (y - y^*)^2}$
    end for
end for
```

With the downsampling factors of $S \times S$, the downsampling ratio can be estimated by $\frac{1}{S\times S} \times 100\%$. Taking $S = 96$ as example, the sparse ratio is $\frac{1}{96\times 96} \times 100\%$ = 0.0108%. It means that only $256 \times 256 \times 0.0108\% \approx 7$ points are kept in a 256×256 original sparse height map H_S. In the following, the sparse height refers to the densified sparse height map **SH**.

2.4. Proxyless Depth-Aware Penalty Neural Architecture Search (PDPNAS) for Unet

Details about the PDPNAS for Unet are shown in Figure 3 and Table 1. Similar to IM2HEIGHT [5], down and upsampling subnetworks are included in the network. Since the representation ability of the network can be improved with the increase of depth, there are 10 mobile inverted residual blocks [39] (MBResblocks), including 5 down MBResblocks and 5 up MBResblocks. Due to the fact that large amounts of parameters in the modern deep learning network lead to overfitting when the train datasets are not large enough, the advantage of MBResblocks is that the total depth of Unet is increased by MBResblocks without including too many parameters. Each block contains 3 convolution layers, including 1 × 1 inverted bottlenecks, k_5xk_5 depth separately convolution, and 1 × 1 channel refinement convolution layers; k_5 can be learned by PDPNAS. Both low-level and high-level features are combined by the skip connections for high-precision height estimation.



(a) Structure of Unet composed of mixed blocks

(b) Down Mixed Block

(c) Up Mixed Block

(d) Down Mobile Inverted Residual Down Mixed Block or Down MBResBlock

(e) Up Mobile Inverted Residual Up Mixed Blocks or Up MBResblock

Figure 3. Network structure of PDNAS for Unet composed of down/up mixed blocks and down/up MBResblocks. Each mixed block contains a set of candidate operations, including 3 × 3, 5 × 5 convolution/transposed convolution, etc.

As illustrated in Figure 3, the Unet with PDPNAS is composed of D mixed blocks instead of the conventional convolution layers with fixed kernel size. Each mixed block contains N candidate operations, including convolution with kernel size 3 × 3, 5 × 5, max/average pooling with kernel size 3 × 3, 5 × 5, identity mapping, etc. Therefore, N parallel paths exist in a single mixed block, and the output of each candidate operations from a mixed node in depth d is denoted as $\mathbf{o}_d \in \mathbb{R}^N, d = 0, 1, 2, \ldots, D$. Different mixed blocks consist of different candidate operations. As shown in Figure 3, convolution or pooling layers are included in down mixed blocks and down MBResblocks, while transposed convolution or upsampling layers are included in the up mixed blocks and up MBResblocks.

Table 1. Details of network structure. D is the total depth of Unet and $D = 17$ in our experiments. i is the # of Down MBResblocks in Unet and $i = 1, 2, 3, 4, 5$. $k_i^\dagger = k_{6-i}$.

	Layer	Kernel Size	Operation	Strides	Depth	Output Size
Input					0	$(256, 256, 3)$
	Down1	$k_1 \times k_1 \times 64$	Conv/Down Pool	2	1	$(128, 128, 64)$
Down	Down2	$k_2 \times k_2 \times 128$	Conv/Pool	2	2	$(64, 64, 128)$
Subnetwork	Down3	$k_3 \times k_3 \times 256$	Conv/Pool	2	3	$(32, 32, 256)$
5×	[Conv1	$1 \times 1 \times 512$	Conv	1		$(32, 32, 512)$
Down	Conv2	$k_i \times k_i \times 512$	Conv	1	3 + i	$(32, 32, 512)$
[MBResblocks]	Conv3]	$1 \times 1 \times 256$	Conv	1		$(32, 32, 256)$
5×	[Conv1	$1 \times 1 \times 512$	Conv	1		$(32, 32, 512)$
Up	Conv2	$k_i^\dagger \times k_i^\dagger \times 512$	Conv	1	D-3-i	$(32, 32, 512)$
[MBResblocks]	Conv3]	$1 \times 1 \times 256$	Conv	1		$(32, 32, 256)$
	Up1	$k_3 \times k_3 \times 256$	match with Down3	2	D-3	$(32, 32, 256)$
Up	Up2	$k_2 \times k_2 \times 128$	match with Down2	2	D-2	$(64, 64, 128)$
Subnetwork	Up3	$k_1 \times k_1 \times 64$	match with Down1	2	D-1	$(128, 128, 64)$
Output	Out	$k_O \times k_O \times 1$	Conv	2	D	$(256, 256, 1)$

In the training stage of PDPNAS, only one of the N parallel paths for a mixed block in depth d is set as active, which is the same as proxylessNAS [28]. It is implemented by the gate parameters $\mathbf{g}_d \in \{0, 1\}^N$, $d = 0, 1, 2, \ldots, D$. In order to make it trainable, the importance parameters $\boldsymbol{\alpha}_d \in \mathbb{R}^N$, $d = 0, 1, 2, \ldots, D$, which are also known as architecture parameters, are introduced to PDPNAS. Therefore, the gate parameters are determined by the architecture parameters and can be denoted as:

$$\mathbf{g}_d = \begin{cases} [1, 0, 0, \ldots, 0], & \text{with a probability of } \alpha_{d,1} \\ [0, 1, 0, \ldots, 0], & \text{with a probability of } \alpha_{d,2} \\ \ldots & \\ [0, 0, 0, \ldots, 1], & \text{with a probability of } \alpha_{d,N} \end{cases} \tag{8}$$

where $d = 0, 1, 2, \ldots, D$. The architecture parameters can be learned by PDPNAS. As illustrated in Figure 3, the active and inactivate paths in Figure 3b–e are plotted as red and black lines, respectively. The output of a mixed node in depth d is denoted as $m_d = \sum_{i=1}^{N} \mathbf{g}_{d,i} \mathbf{o}_{d,i}$.

Due to the existence of skip connections in Unet, a symmetric architecture is required to achieve higher height estimation accuracy. For instance, the activate operation of Down1 layer should be matched with that of the Up3 layer, as described in Table 1. If the active operation of Down1 is a 3 × 3 convolution layer, then that of Up3 should be a 3 × 3 transposed convolution layer. The mismatching between Down1 and Up1 leads to deteriorated performance for height estimation. The gray lines in Figure 3 means that the path of an up mixed block (or down MBResBlocks) will not be selected once the active path of the corresponding down mixed block (or up MBResblocks) has been determined.

In order to further improve the height estimation accuracy of PDPNAS, a depth-aware penalty is proposed to set the proper penalty p for each mixed block according to its depth d and kernel size k. The principal idea of PDPNAS is to solve the problem of sub-optimal searching by PDPNAS for Unet. The Shallower layers prefer smaller kernels according to the results of proxylessNAS [28]. Since they are connected to deeper layers directly by skip connections of Unet, the relative depth of them is changed. In this situation, it is difficult for NAS to find the optimal architecture for Unet. Therefore, the rule-based penalty term we proposed can improve the performance for NAS on Unet. The penalty term p_d is shown in Equation (9) and can be illustrated in Figure 4:

$$p_d = \begin{cases} e^{max(k)/2 \cdot k' \cdot (d/D-0.5)}, & \text{if } k' < 0 \\ e^{-max(k)/2 \cdot k' \cdot ((D-d)/D-0.5)}, & \text{if } k' \geq 0 \end{cases} \tag{9}$$

where k' is the normalized kernel size from -1 to 1. Since k is the kernel size in the set of candidate operations and $k = 1, 3, 5, 7$, k' can be calculated by $k' = k/4 - 1$. d is the depth of the mixed block in Unet. D is the total depth of Unet and $D = 17$ in this paper.

Hence, the architecture parameters α can be iterative optimized as shown in Equation (10):

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\alpha}_{d,i}} = \sum_{j=1}^{N} \frac{\partial \mathcal{L}}{\partial \mathbf{g}_i} \mathbf{p}_i(\delta_{ij} - \mathbf{p}_i) \cdot p_d, \tag{10}$$

where $\mathbf{p}_i$ is the parameters that denote the importance of δ_i in the set candidate operations and $\mathbf{g}_i$ is the binarized gate parameters, as shown in Figure 3. It is involved in the computation graph and can be computed by back-propagation, $\delta_{ij} = 1$ if $i = j$ else $\delta_{ij} = 0$. Details about the architecture parameters $\boldsymbol{\alpha}_i$ are related to proxylessNAS [28].

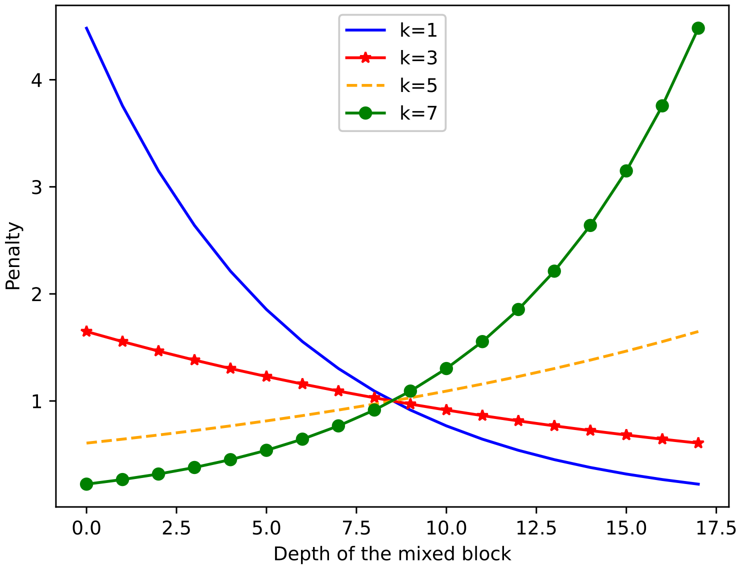


Figure 4. Penalty for a candidate operation set, including convolution and down pooling with a kernel size of 1, 3, 5 and 7.

3. Results and Discussion

3.1. Experiment Setup

SAR images and ground truth DEM are collected over three mountain areas located in Guiyang, Geermu and Huangshan. High-resolution L-band (Advanced Land Observation Satellite-2) ALOS-2 products are used in Guiyang datasets, and DEM 12.5 m is used as the target height map. The resampled spatial resolution of ALOS-2 is 5.7 m in the range direction and 8.5 m in the azimuth direction. A low-resolution (30 m) Shuttle Radar Topography Mission digital elevation model (SRTM) product is used as supplementary sparse input. The elevation range of Guiyang datasets is from 597.70 to 2823.79 m. Images in the Guiyang datasets measure 8130 × 5796 pixels. Both Geermu and Huangshan datasets use C-band Sentinel-1 SAR products and SRTM 30 m for evaluation, which can be easily accessed online. SRTM 90 m is used for supplementary sparse height information. The elevation ranges of Gerrmu and Huangshan datasets are from 2672.74 to 5709.02 m and 100 to 1770.24 m, respectively. Images of Both Geermu and Huangshan datasets measure 6554 × 4800 pixels.

In the Guiyang datasets, the high-resolution L-band SAR images are collected from ALOS-2 products. They are not public and are bought for academic research. The original high-resolution ground truth height maps, which are from DEM 12.5 m, can be downloaded from https://search.asf.alaska.edu/ (accessed on 16 April 2017). They are collected from the L-band PALSAR satellite and generated by InSAR techniques. The supplementary height maps, which are from SRTM 30 m, can be downloaded from https://search.earthdata.nasa.gov/search (accessed on 16 April 2017). They are collected from NASA SRTM C products. In the Geermu and Huangshan datasets, the C-band SAR images can be downloaded from https://asf.alaska.edu/data-sets/sar-data-sets/sentinel-1/sentinel-1-data-and-imagery/ (accessed on 5 September 2022). The method for obtaining ground truth SRTM 30 m is the same as the supplementary SRTM 30 m in Guiyang datasets. The supplementary SRTM 90 m height maps can be downloaded from https://firmware.ardupilot.org/SRTM/ (accessed on 5 September 2022).

The paired SAR images and geocoded height maps in radar coordinates are generated via image registration and geocoding introduced in Section 2.2. As shown in Figure 5, 80% of the images are used as training samples and 20% are used as test samples. Then, all the images (SAR image, height map, sparse height information, distance map) are cut into 256 × 256 small image slices for Unet training. All of the experiments are repeated three times, and both the average and standard deviation are reported for the comparison results.

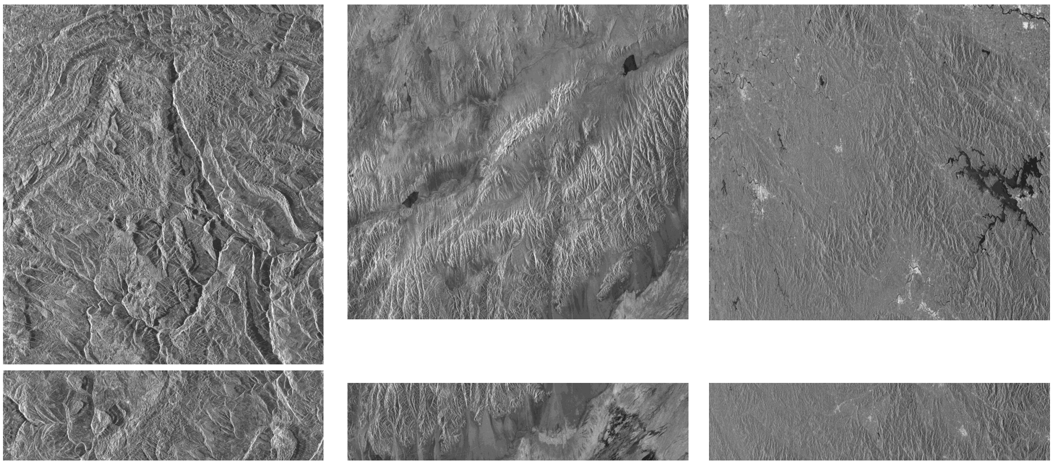


Figure 5. Split train and test data sets for Guiyang, Geermu and Huangshan datasets, respectively. Here, 80% of the images along the azimuth direction are selected for train sets and 20% are selected for test sets. (**a**) is the SAR imagery from Guiyang datasets measuring 8130 × 5796 pixels. (**b**) is the SAR imagery from Geermu datasets measuring 6554 × 4800 pixels. (**c**) is the SAR imagery from Huangshan datasets measuring 6554 × 4800 pixels.

3.2. Evaluation Metrics

To estimate the accuracy of the proposed approaches for height reconstruction from a single SAR image, numerical metrics, which are root mean square error (RMSE) and structural similarity (SSIM) index, are used in our experiments.

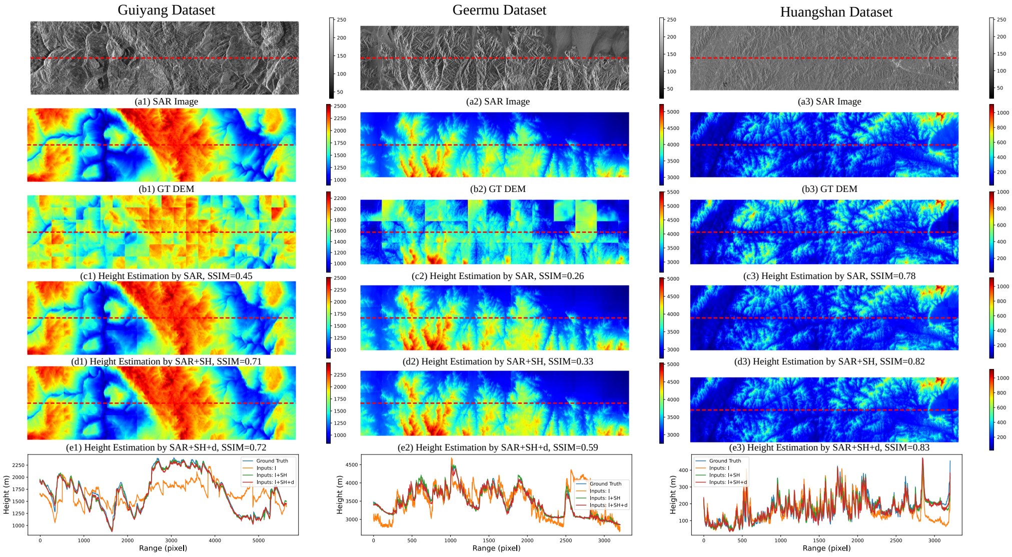
$$\mathrm{RMSE}(\mathbf{x},\mathbf{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{M}\sum_{j=1}^{N}(\mathbf{x}_{ij}-\mathbf{y}_{ij})^2}, \tag{11}$$

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \frac{(2\mu_{\mathbf{x}}\mu_{\mathbf{y}}+C_1)(2\sigma_{\mathbf{xy}}+C_2)}{(\mu_{\mathbf{x}}^2+\mu_{\mathbf{y}}^2+C_1)(\sigma_{\mathbf{x}}^2+\sigma_{\mathbf{y}}^2+C_2)}, \tag{12}$$

where $\mathbf{x}_i$ and $\mathbf{y}_i$ come from the estimated height and height from ground truth, respectively. $\mu_{\mathbf{x}}$ and $\mu_{\mathbf{y}}$ are mean values of $\mathbf{x}$ and $\mathbf{y}$, respectively. $\sigma_{\mathbf{x}}$, $\sigma_{\mathbf{y}}$ and $\sigma_{\mathbf{x}y}$ are standard deviations and cross-covariance for images. $C1$ and $C2$ are small normal numbers. It is capable of comparing local patterns of pixel intensities.

3.3. The Effect of Sparse Height Information and Distance Map

Experiments are carried out for evaluating the effect of sparse height information and PDPNAS for Unet on the three datasets. The height information is estimated from a single SAR image based on a Unet network with various inputs. The results of the whole test sets and an example of local 256 × 256 small slices are shown in Figures 6 and 7. As shown in Figure 6, the height information with SAR is distortions and it is quite different from the ground truth height map. The most possible reason for the errors in height estimation results from the one-to-many relationship between texture and elevation. The results are improved significantly by accepting sparse height information **SH** as inputs. The results of Unet with SAR, sparse height information and distance map as inputs are the best among them. As shown in Figure 7, the shape and texture may be well estimated without sparse height information. However, the estimated height range of it is quite different from the real ground truth height. The estimated height map in some areas outlined in Figure 7 shows the superiority of accepting sparse height information and a distance map as inputs for Unet. It proves that the estimated height with sparse height information and distance map can significantly improve the reliability of the estimated height map. The results in Figure 6(f1–f3) and Figure 7(f1–f3) show that the supplementary inputs **SH**+**d** could significantly improve the performance of height estimation from a single SAR image in mountain areas, especially for that with a wider range of height values.



(a1) SAR Image (a2) SAR Image (a3) SAR Image

(b1) GT DEM (b2) GT DEM (b3) GT DEM

(c1) Height Estimation by SAR, SSIM=0.45 (c2) Height Estimation by SAR, SSIM=0.26 (c3) Height Estimation by SAR, SSIM=0.78

(d1) Height Estimation by SAR+SH, SSIM=0.71 (d2) Height Estimation by SAR+SH, SSIM=0.33 (d3) Height Estimation by SAR+SH, SSIM=0.82

(e1) Height Estimation by SAR+SH+d, SSIM=0.72 (e2) Height Estimation by SAR+SH+d, SSIM=0.59 (e3) Height Estimation by SAR+SH+d, SSIM=0.83

(f1) Estimated Height with SAR vs. SAR+H+D (f2) Estimated Height with SAR vs. SAR+H+D (f3) Estimated Height with SAR vs. SAR+H+D

Figure 6. Height estimation result of a single SAR image based on sparse height information and PDNAS for Unet in three datasets. The downsampling factors of **SH** are 96 × 96. (**a1**–**a3**) are the SAR images, (**b1**–**b3**) are the ground truth height map. (**c1**–**c3**,**d1**–**d3**,**e1**–**e3**) are the estimated height map with SAR, SAR+**SH**, SAR+**SH**+**d** respectively. Elevation lines along the range directions with a fixed middle azimuth value marked by red dash lines in (**a1**–**e1**), (**a2**–**e2**), (**a3**–**e3**) are plotted in (**f1**–**f3**) respectively.

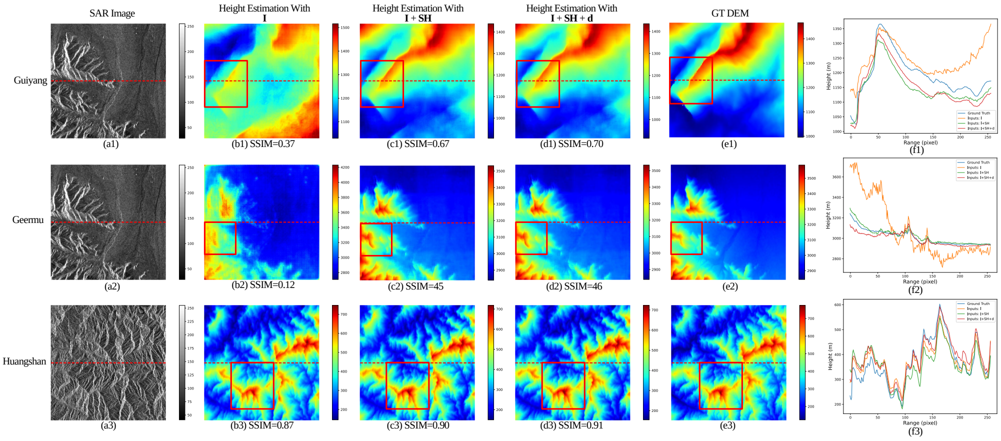


Figure 7. Local height estimation results from a single SAR based on **SH**, **d** and PDPNAS for Unet in Guiyang, Geermu and Huangshan datasets. The samples which measure 256 × 256 pixels are extracted from the whole test set. From left to right, there are (**a1**–**a3**) SAR images **I**, (**b1**–**b3**) estimated height maps with **I** as input, (**c1**–**c3**) estimated height maps with **I** and **SH** as inputs, (**d1**–**d3**) estimated height maps with **I**, **SH** and **d** as inputs. (**e1**–**e3**) are the ground truth height maps. Elevation lines along range directions with a fixed middle azimuth value marked by red dash lines in (**a1**–**e1**), (**a2**–**e2**), (**a3**–**e3**) are plotted in (**f1**–**f3**) respectively. The downsampling factors of **SH** are 96 × 96.

Numerical results on the effect of sparse height information and PDPNAS are shown in Table 2, which are from the Guiyang datasets. The RMSE of an estimated height map with SAR as inputs is 90.30% higher than that with **I**+**SH**+**d** as inputs by observing only 0.0109% height values from low-resolution SRTM products. It is 90.30% higher than that with **I**+**SH**+**d** as inputs and PDPNAS for Unet. In the Geermu datasets, the RMSE is reduced by 95.9% with **I**+**SH** (downsampling factors are 96 × 96)+**d** as inputs and PDPNAS for Unet compared with Unet and single SAR as input. In the Huangshan datasets, the RMSE is reduced by 70.18% with **I**+**SH** (downsampling factors are 96 × 96)+**d** as inputs and PDPNAS for Unet compared with Unet and single SAR as input. Note that the elevation range of Geermu is wider than that of the Guiyang datasets, and the height range of Guiyang is wider than that of the Huangshan datasets. The numerical results suggest that the sparse height information could significantly improve the performance of height estimation, especially for mountain areas with a wide height range.

Other information associated with sparse height information may also affect the height estimation results, as shown in Table 2. In our experiments, we test the effect of including the **d** as the third input besides SAR image and sparse height. The comparison experiments have been carried out when additional sparse height information (distance in our case) is included or not. The experiment results in Table 2 show that the sparse height information can improve the accuracy of single-channel SAR elevation estimation. As listed in Table 2, with lower downsampling factors, the accuracy of the reconstruction result is higher. With a downsampling factors of 32 × 32, the RMSE of height estimation is reduced by about 2.5% when the distance map is included or not, respectively. When downsampling factors are 192 × 192, there is only at most one sparse height point for each 256 × 256 local patch. Even with such sparse height points, it is noted that the estimated height can reach 74.03 m, 65.39 m of RMSE with our model with/without a distance map in the Guiyang datasets. The results proved that our method has achieved remarkable accuracy of height estimation from a single SAR image.

As shown in Table 2, with distance map as the input, all of the RMSE of estimated height is lower than that without it, no matter what downsampling factors are applied. With the increase of sparsity (decrease of percentage points sampled), it has a more positive impact on height estimations. Moreover, when the downsampling ratio of the sparse height information is lower than 0.019%, the SSIM of the estimated height can be improved with the distance information as the third input.

Table 2. Comparison experimental results on the supplementary inputs of sparse height information (SH), distance map (d) and PDPNAS for Unet on three datasets.

Datasets	Downsample Factors	% Points Sampled	Inputs	Include PDPNAS	RMSE (m)	SSIM
Guiyang	—	—	**I**	No	326.74 ± 10.75	0.40 ± 0.05
	32 × 32	0.0976	**SH**	No	30.28 ± 0.25	0.52 ± 1 ×10^{-3}
			I+SH	No	17.94 ± 0.63	0.72 ± 3 ×10^{-3}
			I+SH+d	No	17.49 ± 0.79	0.75 ± 1 ×10^{-3}
			I+SH+d	Yes	**15.70 ± 0.70**	**0.78 ± 7 × 10^{-3}**
	64 × 64	0.0244	**SH**	No	59.02 ± 0.69	0.28 + 2 ×10^{-3}
			I+SH	No	29.45 ± 1.61	0.68 ± 3 ×10^{-3}
			I+SH+d	No	26.65 ± 1.63	0.70 ± 0.03
			I+SH+d	Yes	**24.18 ± 0.28**	**0.74 ± 2 × 10^{-3}**
	96 × 96	0.0109	**SH**	No	78.40 ± 1.54	0.24 ± 4 ×10^{-3}
			I+SH	No	37.84 ± 0.43	0.68 ± 6 ×10^{-3}
			I+SH+d	No	32.95 ± 0.65	0.70 ± 6 ×10^{-3}
			I+SH+d	Yes	**31.70 ± 0.14**	**0.72 ± 5 × 10^{-3}**
	128 × 128	0.0061	**SH**	No	109.91 ± 0.77	0.15 ± 2 ×10^{-3}
			I+SH	No	53.60 ± 0.46	0.64 ± 7 ×10^{-3}
			I+SH+d	No	46.56 ± 0.46	0.64 ± 0.02
			I+SH+d	Yes	**45.01 ± 0.57**	**0.68 ± 7 × 10^{-3}**
	160 × 160	0.0039	**SH**	No	117.72 ± 0.79	0.14 ± 1 ×10^{-3}
			I+SH	No	63.65 ± 0.52	0.63 ± 5 ×10^{-3}
			I+SH+d	No	56.27 ± 0.56	0.63 ± 1 ×10^{-3}
			I+SH+d	Yes	**53.94 ± 0.44**	**0.68 ± 1 × 10^{-3}**
	192 × 192	0.0027	**SH**	No	145.83 ± 1.41	0.11 ± 3 ×10^{-3}
			I+SH	No	74.03 ± 1.296	0.59 + 7 ×10^{-3}
			I+SH+d	No	65.39 ± 0.54	0.62 ± 4 ×10^{-3}
			I+SH+d	Yes	**62.55 ± 0.42**	**0.64 ± 0.08**
Geermu	—	—	**I**	Yes	1015.22 ± 43.63	0.27 ± 0.01
	64 × 64	0.0244	**SH**	Yes	90.96 ± 1.27	0.07 ± 3 ×10^{-3}
			I+SH	Yes	41.61 ± 1.53	0.36 ± 9 ×10^{-3}
			I+SH+d	Yes	**29.77 ± 1.48**	**0.36 ± 6 × 10^{-3}**
	96 × 96	0.0109	**SH**	Yes	77.42 ± 1.74	0.25 ± 4 ×10^{-3}
			I+SH	Yes	43.18 ± 1.25	0.36 ± 9 ×10^{-3}
			I+SH+d	Yes	**41.61 ± 1.53**	**0.37 ± 5 × 10^{-3}**
	128 × 128	0.0061	**SH**	Yes	145.82 ± 2.13	0.03 ± 0.11
			I+SH	Yes	65.44 ± 4.15	0.35 ± 6 ×10^{-3}
			I+SH+d	Yes	**63.93 ± 2.831**	**0.37 ± 2 × 10^{-3}**
Huangshan	—	—	**I**	Yes	124.16 ± 3.18	0.780 ± 5 ×10^{-3}
	64 × 64	0.0244	**SH**	Yes	108.12 ± 0.40	0.06 ± 1 ×10^{-3}
			I+SH	Yes	36.00 ± 1.05	0.81 ± 6 ×10^{-3}
			I+SH+d	Yes	**29.98 ± 2.80**	**0.83 ± 2 × 10^{-3}**
	96 × 96	0.0109	**SH**	Yes	123.82 ± 0.36	0.04 ± 1 ×10^{-3}
			I+SH	Yes	49.06 ± 2.84	0.81 ± 4 ×10^{-3}
			I+SH+d	Yes	**37.02 ± 0.63**	**0.83 ± 2 × 10^{-3}**
	128 × 128	0.0061	**SH**	Yes	161.71 ± 0.58	0.02 ± 2 ×10^{-3}
			I+SH	Yes	60.14 ± 1.03	0.81 ± 0.006
			I+SH+d	Yes	**45.77 ± 1.26**	**0.82 ± 2 × 10^{-3}**

3.4. Comparison of Different Network Structures

Comparison experiments are conducted among six different methods. They are VDSR [18], PackNet [15], IM2HEIGHT [5], 3DMap [10], SAR&phase [6] and ours. The downsampling factors of **SH** in this experiment are 96 × 96. The comparison experiments are performed on the Guiyang datasets, and the numerical results are listed in Table 3. It shows that our model has the best performance among all these six network structures. The RMSE of IM2HEIGTH together with **SH+d** as inputs is 40.06% higher than our methods. It shows the superiority of our network, which uses an advanced Unet and PDPNAS. Since the structure of Unet is not reported in 3DMap [10], it is set to be the same as ours for comparison in this paper. The RMSE of 3D MAP Reconstruction together with **SH+d** as inputs is 14.72% higher than ours, and it suggests that the building adaptive loss function is not suitable for height estimation in mountain areas. The RMSE of SAR&phase is 91.92% higher than ours, while that of SAR&phase together with the same structure of Unet to us, using PDPNAS for Unet and **SH+d** as inputs, is 5.32% higher than ours. It suggests that the meaningless phase map is not helpful for single SAR imagery height estimation.

As illustrated in Figure 8, the results of ours are more realistic than the others. The results of IM2HEIGHT are worse than 3D MAP Reconstruction, SAR&phase together with PDPNAS for Unet and **SH+d** as inputs, and ours. As shown in Figure 9, the details of our estimated height map are more realistic, and they are more similar to that of the ground truth height map than the others, especially for the outline areas. The results in Figure 9(g1–g6) show that the estimated height map of ours has lower reconstructed errors compared with other methods.

The RMSE and SSIM of IM2HEIGHT, 3DMap, SAR&phase and ours are illustrated in Figure 10. In total, 132 unique 256 × 256 local patches of the test scene are used for evaluation. The RMSE of our methods is around 30 m, and it provides a relatively high accuracy for height estimation compared with other methods. Considering that few percentages of height points from low-resolution SRTM are used for height estimation, our method has remarkable performance on height estimation with a single SAR image.

Table 3. Comparison experiments between different methods in Guiyang datasets. The sparse factors of **SH** are 96 × 96.

Method	**Inputs**	**Structure of Unet Same to Us**	**Include PDPNAS**	**RMSE (m)**	**SSIM**
VDSR [18]	**I+SH+d**	No	No	79.36 ± 6.92	$0.31 \pm 6 \times 10^{-3}$
PackNet [15]	**I+SH+d**	No	No	59.55 ± 1.80	$0.49 \pm 2 \times 10^{-3}$
IM2HEIGHT [5]	**I+SH+d**	No	No	52.89 ± 0.37	$0.31 \pm 5 \times 10^{-3}$
3DMap [10]	**I+SH+d**	Yes	No	37.17 ± 1.11	0.54 ± 0.01
SAR&phase [6]	**I+phase**	No	No	392.22 ± 3.68	$0.01 \pm 3 \times 10^{-3}$
	I+SH+d+phase	Yes	No	33.48 ± 0.25	$0.72 \pm 5 \times 10^{-3}$
ours	**I+SH+d**	Yes	No	32.95 ± 0.65	$0.70 \pm 6 \times 10^{-3}$
		Yes	Yes	**31.70 ± 0.14**	$\mathbf{0.73 \pm 5 \times 10^{-3}}$

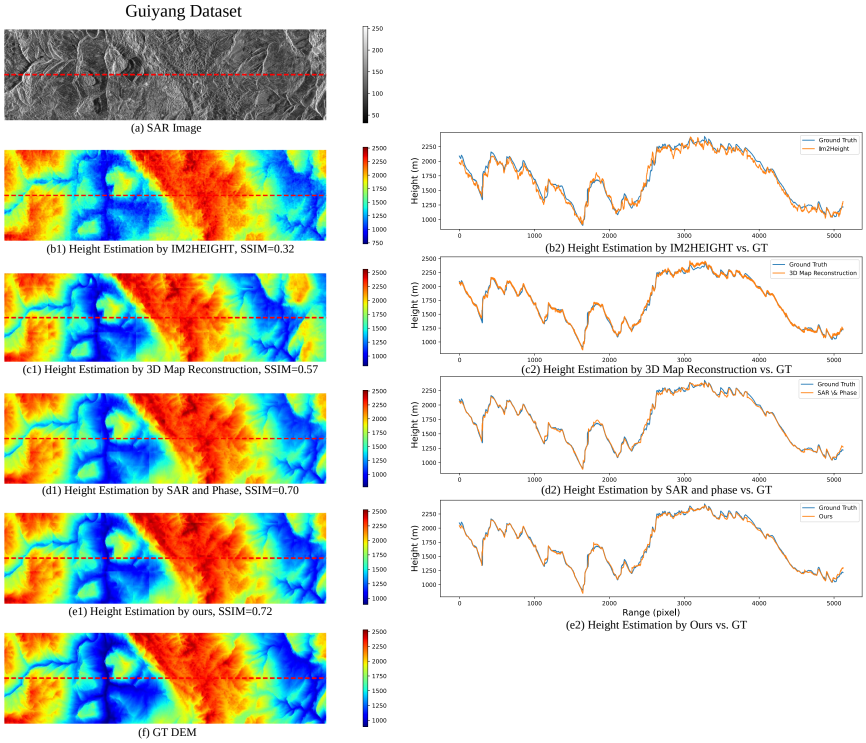


Figure 8. Height estimation result of different methods in Guiyang datasets. From top to bottom, (**a**) is SAR image. (**b1**–**e1**) are the estimated height by IM2HEIGHT, 3D map reconstruction, SAR&phase and ours, respectively. Elevation lines along range directions with a fixed middle azimuth value marked by red dash lines in (**b1**–**e1**) are plotted in (**b2**–**e2**) respectively. (**f**) is the ground truth height map. The downsampling factors of **SH** are 96 × 96.

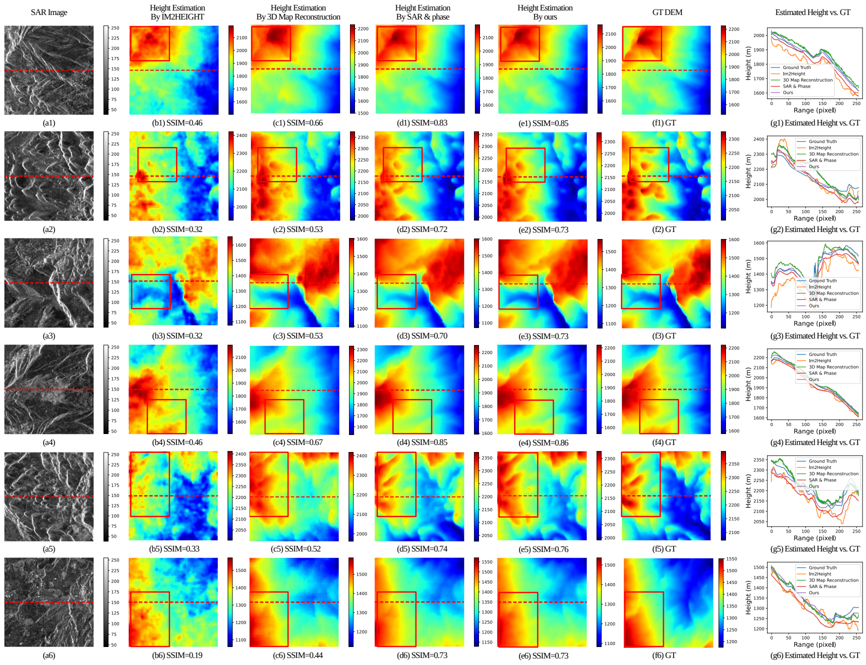


Figure 9. Six local height patches estimated by different methods in Guiyang datasets. From the left to right, (**a1**–**a6**) are the SAR images. (**b1**–**b6**,**c1**–**c6**,**d1**–**d6**,**e1**–**e6**) are the estimated height map by IM2HEIGHT, 3D Map Reconstruction, SAR&phase and our method, respectively. (**f1**–**f6**) are the ground truth height maps. Elevation lines along range directions with a fixed middle azimuth value marked by red dash lines in (**a1**–**f1**), (**a2**–**f2**), (**a3**–**f3**), (**a4**–**f4**), (**a5**–**f5**), (**a6**–**f6**) are plotted in (**g1**–**g6**) respectively. The downsampling factors of **SH** are 96 × 96.

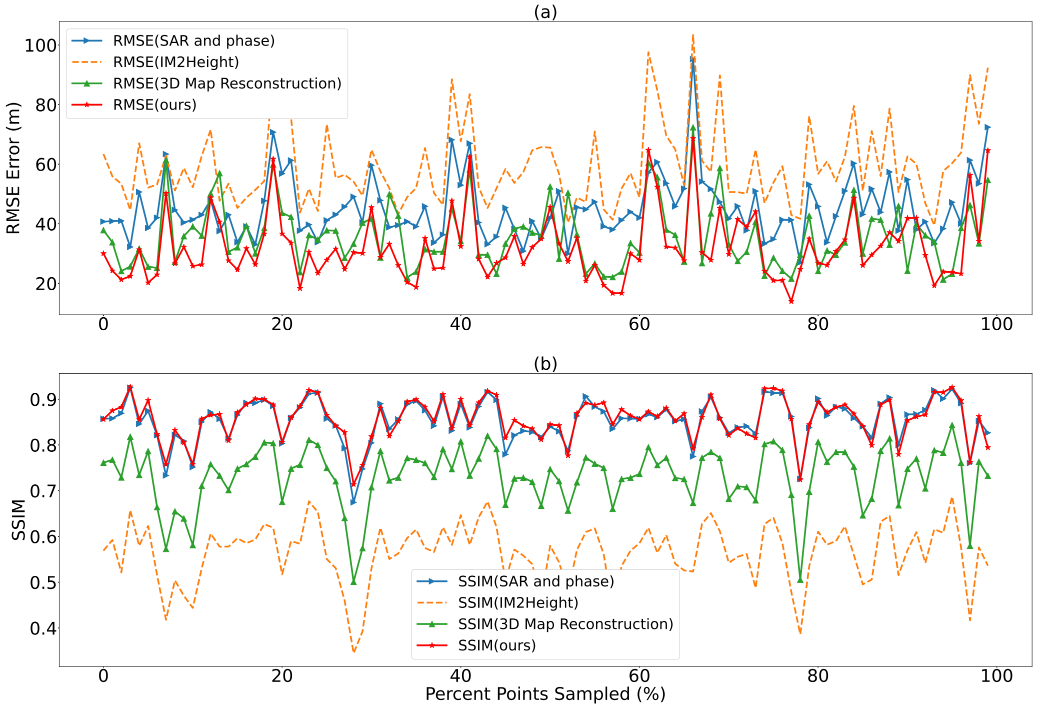


Figure 10. (**a**) RMSE and (**b**) SSIM of SAR&phase, IM2HEIGHT, 3D map reconstruction and ours on 132 unique 256 × 256 local patches of the test scene with downsampling factors of 96 × 96.

3.5. The Effect of PDPNAS

Comparison experiments on the effect of PDPNAS have been conducted in Guiyang datasets, and the results are shown in Table 2 and Figure 11. As shown in Table 2, the RMSE is reduced by about 3% in all the cases. It suggests that PDPNAS could find optimal architectures for Unet, and it is helpful for singe SAR height estimation. The result in Figure 11 is consistent with that of Table 2. It is worth noting that the SSIM is increased significantly, as shown in Figure 11, and it suggests that the estimated height map with PDPNAS is more realistic that that without PDPNAS. Although the improvement is not that outstanding compared with sparse height information, the work of PDNAS is still valuable in that it makes it possible for Unet to be deployed on a hard device.

Comparison experiments among Unet, NAS-Unet and PDPNAS for Unet are conducted, and the numerical results are shown in Table 4. The RMSE of NAS-Unet is increased by 1.76% compared with symmetric Unet. It suggests that the asymmetric architecture searched by NAS-Unet leads to a deteriorated performance compared with Unet. The RMSE of PDPNAS for Unet with symmetric constraint and depth-aware penalty is reduced by 3.79% and 5.49% compared with Unet and NAS-Unet, respectively. It shows the superiority of PDPNAS for height estimation with a single SAR imagery. Comparison experiments have also been conducted on the effect of MBResblocks, and the results are shown in Table 4. The RMSE of PDNAS without MBResblocks is 40.06% higher than that with MBResblocks. It shows the superiority of MBResblocks for Unet.

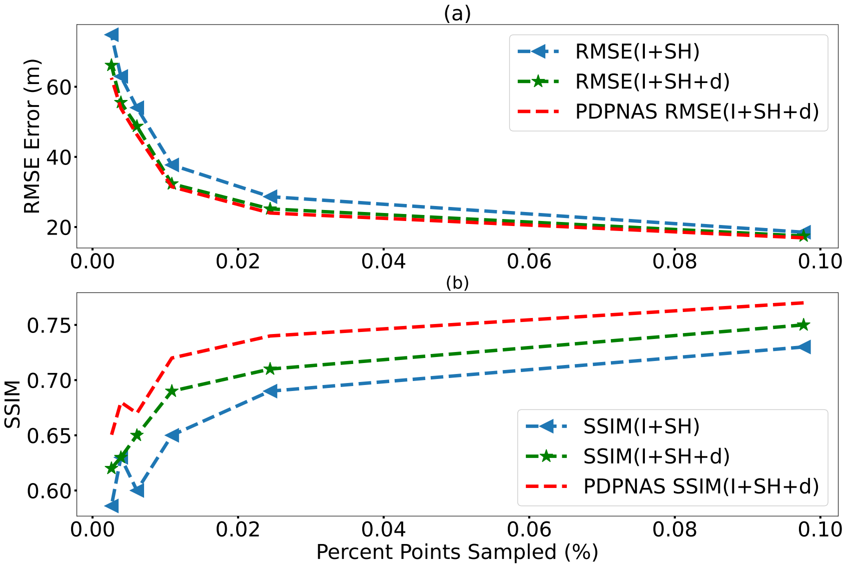


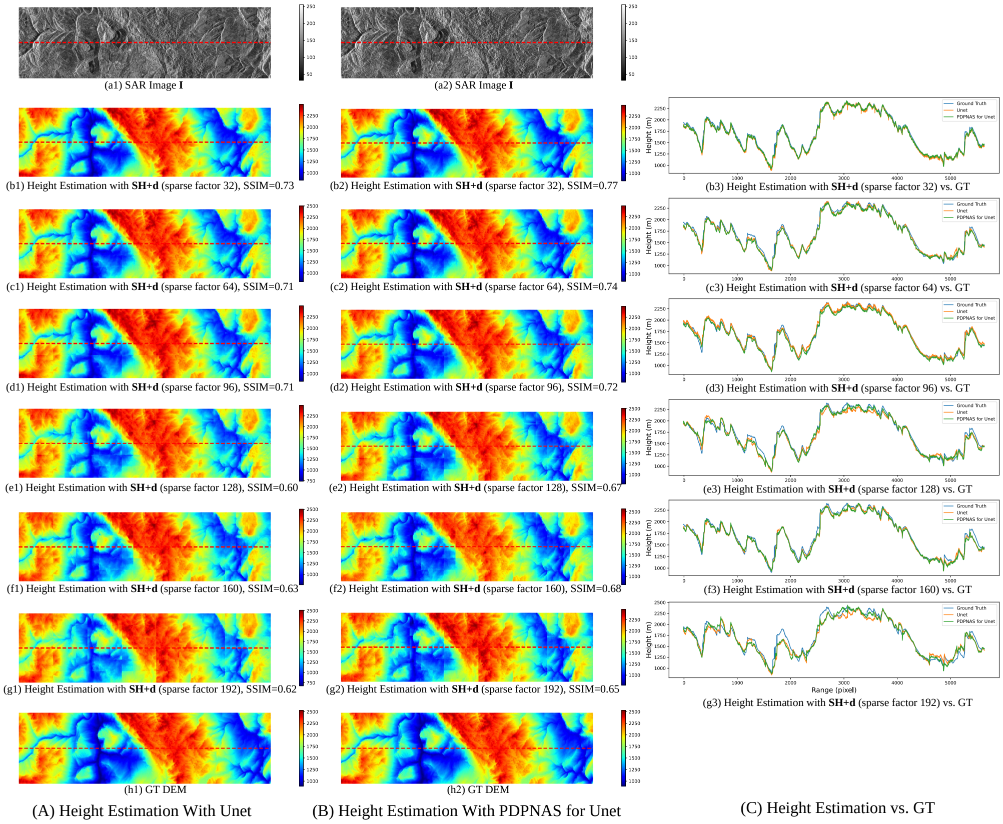
Figure 11. RMSE and SSIM of Unet with **I+SH** and **I+SH+d** and PDPNAS for Unet with **I+SH+d** on Guiyang datasets. x-axis represents the downsampling ratios (or the percentages of sampled points) of **SH**. (**a**) The RMSE results. (**b**) The SSIM results.

Table 4. Comparison experiments on the effect of PDPNAS for Unet in Guiyang datasets. The downsampling factors of sparse height information are 96 × 96.

Methods	**Include MBResblocks**	**Is Symmteric**	**Includes Depth-Aware Penalty**	**RMSE (m)**	**SSIM**
Unet	Yes	Yes	No	32.95 ± 0.65	$0.70 \pm 6 \times 10^{-3}$
NAS-Unet	Yes	No	No	33.54 ± 0.63	0.69 ± 0.01
PDPNAS for Unet	No	Yes	Yes	52.89 ± 0.37	$0.31 \pm 5 \times 10^{-3}$
	Yes	No	Yes	32.67 ± 0.34	$0.71 \pm 3 \times 10^{-3}$
	Yes	Yes	Yes	**31.70 ± 0.14**	$\mathbf{0.72 \pm 5 \times 10^{-3}}$

3.6. Comparison on the Effect of Sparse Height with Various Sparsity

Comparison experiments are carried out to evaluate the effect of sparse height with various sparsity during height reconstruction in the three datasets. In the Guiyang dataset, the downsampling factors of sparse height information are 32 × 32, 64 × 64, 96 × 96, 128 × 128, 160 × 160, and 192 × 192. In the Geermu and Huangshan datasets, the downsampling factors of sparse height information are 64 × 64, 96 × 96, 128 × 128. The height estimation results are illustrated in Figures 12 and 13. As shown in Figure 12, the estimated height map with lower downsampling factors is nearly the same as the ground truth height map. Even with an extremely sparse height as input (downsampling factors are 192 × 192), the estimated height map is realistic and similar to the ground truth height map. The results on PDPNAS are consistent with those without PDPNAS for Unet.



(a1) SAR Image **I**

(a2) SAR Image **I**

(b1) Height Estimation with **SH+d** (sparse factor 32), SSIM=0.73

(b2) Height Estimation with **SH+d** (sparse factor 32), SSIM=0.77

(b3) Height Estimation with **SH+d** (sparse factor 32) vs. GT

(c1) Height Estimation with **SH+d** (sparse factor 64), SSIM=0.71

(c2) Height Estimation with **SH+d** (sparse factor 64), SSIM=0.74

(c3) Height Estimation with **SH+d** (sparse factor 64) vs. GT

(d1) Height Estimation with **SH+d** (sparse factor 96), SSIM=0.71

(d2) Height Estimation with **SH+d** (sparse factor 96), SSIM=0.72

(d3) Height Estimation with **SH+d** (sparse factor 96) vs. GT

(e1) Height Estimation with **SH+d** (sparse factor 128), SSIM=0.60

(e2) Height Estimation with **SH+d** (sparse factor 128), SSIM=0.67

(e3) Height Estimation with **SH+d** (sparse factor 128) vs. GT

(f1) Height Estimation with **SH+d** (sparse factor 160), SSIM=0.63

(f2) Height Estimation with **SH+d** (sparse factor 160), SSIM=0.68

(f3) Height Estimation with **SH+d** (sparse factor 160) vs. GT

(g1) Height Estimation with **SH+d** (sparse factor 192), SSIM=0.62

(g2) Height Estimation with **SH+d** (sparse factor 192), SSIM=0.65

(g3) Height Estimation with **SH+d** (sparse factor 192) vs. GT

(h1) GT DEM

(h2) GT DEM

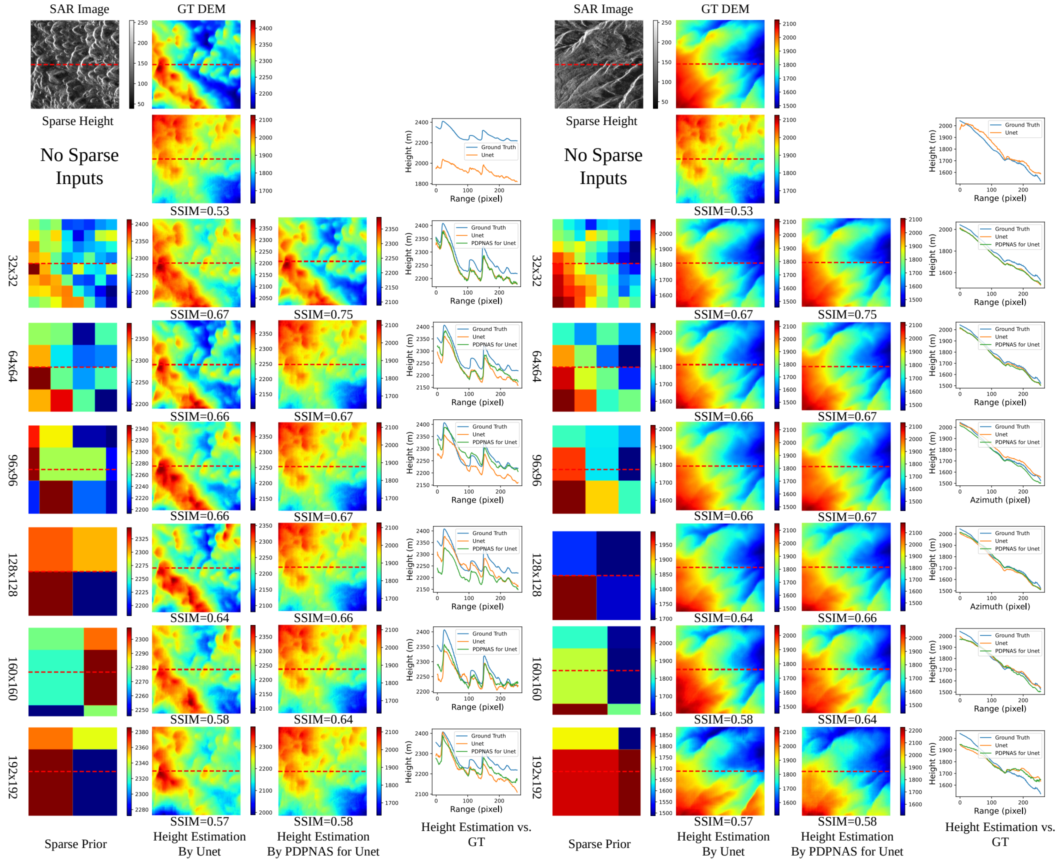
(A) Height Estimation With Unet

(B) Height Estimation With PDPNAS for Unet

(C) Height Estimation vs. GT

Figure 12. Comparison experimental results of different sparse height information as inputs with/without PDPNAS for Unet in Guiyang datasets. Different sparse height information was applied, and the downsampling factors are 32 × 32, 64 × 64, 96 × 96, 128 × 128, 160 × 160, and 192 × 192. Elevation lines along range directions with a fixed middle azimuth value marked by red dash lines in (**a1**–**g1**), (**a2**–**g2**) are plotted in (**b3**–**g3**) respectively.

As shown in Figure 13, two examples are selected from Guiyang datasets to show the effect of sparse height information with different downsampling factors. The first row presents the SAR image and the corresponding ground truth DEM. Reconstruction results without sparse height information as input are presented in the second row. In another word, it means that the height is estimated by a SAR image alone with the Unet network. The results show that the estimated height of some areas is in a reasonable range. However, there is a large disparity between the reconstructed height and the ground truth DEM in some areas. The denser the sparse height information is, the better the details of the estimated height maps are, and the closer they are to the ground truth DEM. It is a trade-off between the quality of estimated height map and the resolution of supplementary low-resolution SRTM products. It is worth noting that even in the case of downsampling of 192 × 192, our model still performances very well, and it presents a clear texture and looks similar to ground truth DEM. It means that only one height point is observed in a 256 × 256 image slice. The quantitative results are shown in Table 2, which is consistent with the above results.



(**a**) Sample 1 (**b**) Sample 2

Figure 13. Two local estimated height maps with various sparsity in Guiyang datasets. Comparison experiments between Unet with and without PDPNAS are also shown in the figure. The down-sampling factors are 32 × 32, 64 × 64, 96 × 96, 128 × 128, 160 × 160, and 192 × 192, respectively. Elevation lines along range directions with a fixed middle azimuth value marked by red dash lines in the figures are plotted in 4th and 8th columns respectively.

3.7. Layover and Shadow

Comparison experiments are performed to evaluate the impact of layover and shadow areas on the three datasets. As illustrated in Figures 14 and 15, the areas of layover and shadow are marked as red in both SAR and height maps of the Guiyang, Geermu and Huangshan datasets. Since the elevation ranges of Geermu are wider and the maximum elevation is larger than the other datasets, layover and shadow occur more often compared with the Guiyang and Huangshan datasets, which is consistent with the description in Section 2.1. Meanwhile, the areas of layover are larger than those of shadow in the Geermu datasets. By contrast, the maximum elevation in the Huangshan datasets is lower and the elevation ranges are shallower than those in the other two datasets, and fewer layover/shadow areas exist in the Huangshan datasets.

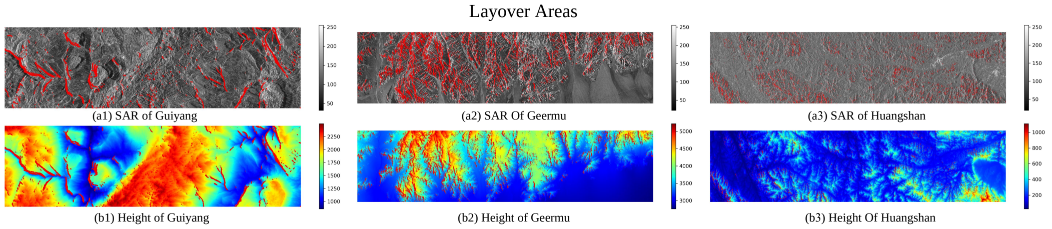


Figure 14. Layover areas are marked as red points in three datasets.

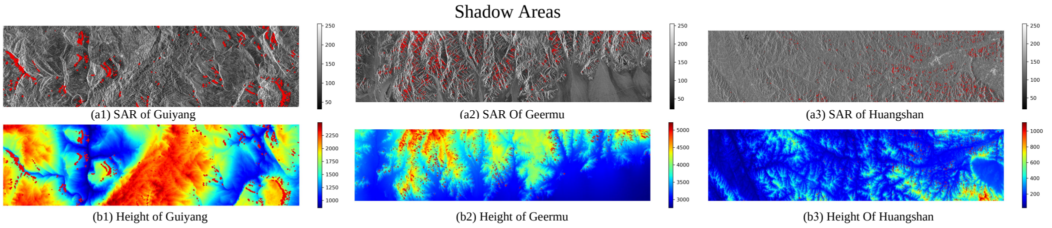


Figure 15. Shadow areas are marked as red points in three datasets.

Since the layover and shadow areas are inconsistent as shown in Figures 14 and 15, the sliding window cannot be used to calculate SSIM. Hence, SSIM is not suitable for elevation in layover and shadow areas. Therefore, mean absolute relative error (MARE) is added as an elevation metric besides RMSE in this section. The MARE can be calculated as following:

$$\text{MARE}(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{|\mathbf{x}_i - \mathbf{y}_i|}{max|\mathbf{y}|} \right), \tag{13}$$

where $\mathbf{x}_i$ and $\mathbf{y}_i$ come from the estimated and ground truth height maps, respectively.

As listed in Table 5, comparison experiments are conducted with different **SH**. In the Guiyang datasets, our model performs well in shadow areas. When the downsampling factors of **SH** are lower than 96 × 96 (which can be 32 × 32, 64 × 64, 128 × 128), both the RMSE and MARE of the estimated height in the shadow areas are nearly the same as the areas without layover and shadow. It is worth noting that it leads to deteriorating performance in the areas of layover; both the RMSE and MARE in layover areas are nearly 3× higher compared with the other areas. When the layover and shadow areas are removed from the test sets, the RMSE of the estimated height map is reduced to about 30.09 m.

In the Geermu and Huangshan datasets, the RMSE of the areas with layover and shadow areas is lower than that without them, which is consistent with the results in the Guiyang datasets. The RMSE of Geermu is higher than the other datasets with the same **SH** as input, even without layover and shadow. The first reason is that the maximum height of the Geermu dataset is larger than the others. Since the output height is normalized to [0, 1], the same error from the estimated height of the three datasets leads to tremendous distinction for RMSE. By contrast, the RMSE of the Huangshan datasets with the same **SH** is lower than the other two datasets; for that, the maximum height of the Huangshan datasets is lower than the others. The second reason for the higher RMSE in the Geermu dataset is that the areas of layover and shadows are larger than those of the other datasets, as shown in Figures 14 and 15. It suggests that the geometric distortions caused by the side-looking imaging geometry of SAR have a great negative impact on the estimated height map. Therefore, future work is to ease the negative effect of layover on single SAR height estimation.

Note that the mean absolute relative error (MARE) of the Huangshan datasets is higher than that of the other two datasets, even with a lower RMSE. It suggests that a more intelligent method is required to generate a normalized height map. Moreover, the RMSE of the Guiyang datasets is lower than that of the other two datasets with the same **SH** as inputs. Compared with the results of the Huangshan datasets, the larger areas of layover and shadow in the Guiyang datasets do not leads to higher RMSE in the estimated height map. The reason is that the spatial resolution of SAR applied in the Guiyang datasets is higher than the others, and it has a significant positive impact on the estimated height map. Therefore, a high-resolution SAR product is required to generate a high precision estimated height map.

Table 5. Comparison experiments on the layover and shadow areas on the three datasets.

Datasets	Downsample Factors	% Points Sampled	Include Layover	Include Shadow	Include Others	RMSE (m)	MARE (%)
Guiyang	32 × 32	0.0976	Yes	Yes	Yes	15.70 ± 0.70	0.54 ± 3 × 10^{-3}
			Yes	No	No	72.73 ± 1.41	3.47 ± 0.01
			No	Yes	No	15.87 ± 0.63	0.66 ± 9 × 10^{-3}
			No	No	Yes	**12.39 ± 0.19**	**0.50 ± 4 × $\mathbf{10^{-3}}$**
	64 × 64	0.0244	Yes	Yes	Yes	23.38 ± 0.28	1.00 ± 2 × 10^{-3}
			Yes	No	No	81.71 ± 1.32	4.14 ± 0.01
			No	Yes	No	24.60 ± 0.62	1.17 ± 2 × 10^{-3}
			No	No	Yes	**21.52 ± 0.39**	**0.95 ± 7 × $\mathbf{10^{-3}}$**
	96 × 96	0.0109	Yes	Yes	Yes	31.70 ± 0.14	1.38 ± 2 × 10^{-3}
			Yes	No	No	93.25 ± 1.46	4.78 ± 0.02
			No	Yes	No	31.66 ± 0.65	1.45 ± 3 × 10^{-3}
			No	No	Yes	**30.09 ± 0.18**	**1.33 ± 8 × $\mathbf{10^{-3}}$**
	128 × 128	0.0061	Yes	Yes	Yes	45.01 ± 0.57	2.01 ± 4 × 10^{-3}
			Yes	No	No	104.45 ± 1.60	5.62 ± 0.01
			No	Yes	No	56.30 ± 0.55	2.54 ± 6 × 10^{-3}
			No	No	Yes	**43.11 ± 0.35**	**1.96 ± 3 × $\mathbf{10^{-3}}$**
	160 × 160	0.0039	Yes	Yes	Yes	53.94 ± 0.44	2.38 ± 1 × 10^{-3}
			Yes	No	No	119.69 ± 1.45	6.53 ± 0.01
			No	Yes	No	57.91 ± 0.88	2.67 ± 1 × 10^{-3}
			No	No	Yes	**50.52 ± 0.76**	**2.32 ± 1 × $\mathbf{10^{-3}}$**
	192 × 192	0.0027	Yes	Yes	Yes	62.55 ± 0.42	2.88 ± 2 × 10^{-3}
			Yes	No	No	114.76 ± 1.87	6.19 ± 0.02
			No	Yes	No	65.64 ± 1.07	3.18 ± 5 × 10^{-3}
			No	No	Yes	**60.96 ± 0.80**	**2.84 ± 1 × $\mathbf{10^{-3}}$**
Geermu	64 × 64	0.0244	Yes	Yes	Yes	29.77 ± 1.48	0.46 ± 1 × 10^{-3}
			Yes	No	No	67.82 ± 1.58	1.28 ± 0.03
			No	Yes	No	37.62 ± 1.16	0.72 ± 0.01
			No	No	Yes	**25.06 ± 0.23**	**0.45 ± 1 × $\mathbf{10^{-3}}$**
	96 × 96	0.0109	Yes	Yes	Yes	41.61 ± 1.53	0.72 ± 1 × 10^{-3}
			Yes	No	No	87.17 ± 1.40	1.69 ± 0.01
			No	Yes	No	57.58 ± 1.10	1.15 ± 4 × 10^{-3}
			No	No	Yes	**39.16 ± 0.61**	**0.70 ± 1 × $\mathbf{10^{-3}}$**
	128 × 128	0.0061	Yes	Yes	Yes	63.93 ± 2.83	1.01 ± 1 × 10^{-3}
			Yes	No	No	125.83 ± 1.39	2.54 ± 0.01
			No	Yes	No	101.13 ± 1.51	2.10 ± 4 × 10^{-3}
			No	No	Yes	**55.90 ± 0.26**	**0.99 ± 1 × $\mathbf{10^{-3}}$**
Huangshan	64 × 64	0.0244	Yes	Yes	Yes	29.98 ± 2.80	5.58 ± 0.06
			Yes	No	No	43.90 ± 1.36	6.26 ± 0.03
			No	Yes	No	34.74 ± 1.51	6.23 ± 0.01
			No	No	Yes	**23.23 ± 0.16**	**3.92 ± 0.02**
	96 × 96	0.0109	Yes	Yes	Yes	37.02 ± 0.63	7.85 ± 0.02
			Yes	No	No	49.25 ± 1.44	8.91 ± 0.02
			No	Yes	No	47.35 ± 1.36	8.87 ± 0.04
			No	No	Yes	**33.77 ± 0.44**	**6.76 ± 0.01**
	128 × 128	0.0061	Yes	Yes	Yes	45.77 ± 1.26	9.25 ± 0.10
			Yes	No	No	59.04 ± 1.74	10.83 ± 0.80
			No	Yes	No	58.82 ± 0.31	10.78 ± 0.30
			No	No	Yes	**42.69 ± 0.66**	**8.00 ± 0.09**

4. Conclusions

In this paper, we propose a single SAR imagery height estimation method via PDPNAS for Unet with the help of a sparse height **SH** and distance map **d** in mountain areas. Even though the ground truth (GT) points in **SH** are extremely sparse (for instance, only one GT point in a 256 × 256 small slices), it could reduce the search space and improve the accuracy of height estimation by 80.86% compared with only accepting SAR as input. The performance of Unet for height estimation is improved significantly in mountain areas. Then, flexible downsampling methods are proposed to generate **SH** with different sparse ratios; thus, it could be convenient to produce paired SAR and **SH** for Unet training. In order to further improve the accuracy of height estimation, a PDPNAS method is proposed for Unet; skip connections and mobile inverted residual blocks are included in the Unet. PDPNAS tackles the problem of sub-optimal searching for proxylessNAS on Unet via a depth-aware penalty term. The accuracy of Unet is improved by 3% via PDPNAS in the Guiyang datasets. Several experiments on three datasets are carried out to test the effects of sparse height information and PDPNAS. All the experiments have demonstrated that our model can generate height information with high quality from a single SAR image, which could be helpful for the research of height estimation from both hardware and software perspectives. It will be helpful for the development of new methods for fast worldwide DEM mapping as well. In the future, we will provide the extensive experiments of height estimation with SAR and 3D LiDAR point clouds as inputs. Since that the magic of sparse height has been proved in this paper even with an extremely high sparsity, it makes it possible to obtain a high-resolution estimated height map with a single SAR imagery and low-cost LiDAR products as inputs.

Author Contributions: Conceptualization, M.X. and Q.L.; methodology, M.X.; software, M.X.; validation, M.X.; formal analysis, M.X.; investigation, M.X.; resources, Q.L., Z.Z.; data curation, M.X.; writing—original draft preparation, M.X. and Q.L.; writing—review and editing, M.X. and Q.L.; visualization, M.X.; supervision, Q.L.; project administration, J.L.; funding acquisition, Q.L. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Key Project of Tianjin Natural Science Foundation (21JCZDJC00670), by the National Engineering Laboratory for Digital Construction and Evaluation Technology of Urban Rail Transit (grant No. 2021ZH04), and by the Tianjin Transportation Science and Technology Development Project (grant No. 2020-02, 2022-40), and the National Natural Science Foundation of China Grant (grant No. 41601446). Thanks Ms. Haiying Gao and Tongtong Wang for the support of SAR data processing.

Data Availability Statement: Not applicable.

Acknowledgments: Thanks Haiying Gao and Tongtong Wang for the support of SAR data processing.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Rizzoli, P.; Martone, M.; Gonzalez, C.; Wecklich, C.; Tridon, D.B.; Bräutigam, B.; Bachmann, M.; Schulze, D.; Fritz, T.; Huber, M.; et al. Generation and performance assessment of the global TanDEM-X digital elevation model. *ISPRS J. Photogramm. Remote Sens.* **2017**, *132*, 119–139. [CrossRef]
2. Hoiem, D.; Efros, A.A.; Hebert, M. Recovering surface layout from an image. *Int. J. Comput. Vis.* **2007**, *75*, 151–172. [CrossRef]
3. Saxena, A.; Sun, M.; Ng, A.Y. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 824–840. [CrossRef] [PubMed]
4. Karsch, K.; Liu, C.; Kang, S.B. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 2144–2158. [CrossRef] [PubMed]
5. Mou, L.; Zhu, X.X. IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network. *arXiv* **2018**, arXiv:1802.10249.

6. Costante, G.; Ciarfuglia, T.A.; Biondi, F. Towards monocular digital elevation model (DEM) estimation by convolutional neural networks-Application on synthetic aperture radar images. In Proceedings of the EUSAR 2018, 12th European Conference on Synthetic Aperture Radar, Aachen, Germany, 4–7 June 2018; VDE: Berlin, Germany , 2018; pp. 1–6.
7. El-Darymli, K.; McGuire, P.; Power, D.; Moloney, C. Rethinking the phase in single-channel SAR imagery. In Proceedings of the 2013 14th International Radar Symposium (IRS), Dresden, Germany, 19-21 June 2013; IEEE: Piscataway, NJ, USA, 2013; Volume 1, pp. 429–436.
8. Amirkolaee, H.A.; Arefi, H. Height estimation from single aerial images using a deep convolutional encoder-decoder network. *ISPRS J. Photogramm. Remote Sens.* **2019**, *149*, 50–66. [CrossRef]
9. Ghamisi, P.; Yokoya, N. Img2dsm: Height simulation from single imagery using conditional generative adversarial net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 794–798. [CrossRef]
10. Son, C.; Park, S.Y. 3D Map Reconstruction From Single Satellite Image Using a Deep Monocular Depth Network. In Proceedings of the 2022 Thirteenth International Conference on Ubiquitous and Future Networks (ICUFN), Barcelona, Spain, 5–8 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 5–7.
11. Carvalho, M.; Le Saux, B.; Trouvé-Peloux, P.; Champagnat, F.; Almansa, A. Multitask learning of height and semantics from aerial images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 1391–1395. [CrossRef]
12. Srivastava, S.; Volpi, M.; Tuia, D. Joint height estimation and semantic labeling of monocular aerial images with CNNs. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 5173–5176.
13. Zamir, A.R.; Sax, A.; Shen, W.; Guibas, L.J.; Malik, J.; Savarese, S. Taskonomy: Disentangling task transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3712–3722.
14. Mahmud, J.; Price, T.; Bapat, A.; Frahm, J.M. Boundary-aware 3D building reconstruction from a single overhead image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 441–451.
15. Mallya, A.; Lazebnik, S. Packnet: Adding multiple tasks to a single network by iterative pruning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7765–7773.
16. Liu, C.J.; Krylov, V.A.; Kane, P.; Kavanagh, G.; Dahyot, R. IM2ELEVATION: Building height estimation from single-view aerial imagery. *Remote Sens.* **2020**, *12*, 2719. [CrossRef]
17. Amini Amirkolaee, H.; Arefi, H. Generating a highly detailed DSM from a single high-resolution satellite image and an SRTM elevation model. *Remote Sens. Lett.* **2021**, *12*, 335–344. [CrossRef]
18. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
19. Xia, Z.; Sullivan, P.; Chakrabarti, A. Generating and exploiting probabilistic monocular depth estimates. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 65–74.
20. Pravitasari, A.A.; Iriawan, N.; Almuhayar, M.; Azmi, T.; Irhamah, I.; Fithriasari, K.; Purnami, S.W.; Ferriastuti, W. UNet-VGG16 with transfer learning for MRI-based brain tumor segmentation. *TELKOMNIKA (Telecommun. Comput. Electron. Control)* **2020**, *18*, 1310–1318. [CrossRef]
21. Pellegrin, L.; Martinez-Carranza, J. Towards depth estimation in a single aerial image. *Int. J. Remote Sens.* **2020**, *41*, 1970–1985. [CrossRef]
22. Stan, S.; Rostami, M. Unsupervised model adaptation for continual semantic segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 2593–2601.
23. Wibowo, A.; Triadyaksa, P.; Sugiharto, A.; Sarwoko, E.A.; Nugroho, F.A.; Arai, H.; Kawakubo, M. Cardiac Disease Classification Using Two-Dimensional Thickness and Few-Shot Learning Based on Magnetic Resonance Imaging Image Segmentation. *J. Imaging* **2022**, *8*, 194. [CrossRef]
24. Simon, C.; Koniusz, P.; Nock, R.; Harandi, M. Adaptive subspaces for few-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 4136–4145.
25. Zhang, J.; Xing, M.; Sun, G.C.; Shi, X. Vehicle Trace Detection in Two-Pass SAR Coherent Change Detection Images With Spatial Feature Enhanced Unet and Adaptive Augmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
26. Bender, G.; Kindermans, P.J.; Zoph, B.; Vasudevan, V.; Le, Q. Understanding and simplifying one-shot architecture search. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; PMLR: London, UK, 2018; pp. 550–559.
27. Liu, H.; Simonyan, K.; Yang, Y. Darts: Differentiable architecture search. *arXiv* **2018**, arXiv:1806.09055.
28. Cai, H.; Zhu, L.; Han, S. ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
29. Weng, Y.; Zhou, T.; Li, Y.; Qiu, X. Nas-unet: Neural architecture search for medical image segmentation. *IEEE Access* **2019**, *7*, 44247–44257. [CrossRef]
30. Park, Y.; Guldmann, J.M. Creating 3D city models with building footprints and LIDAR point cloud classification: A machine learning approach. *Comput. Environ. Urban Syst.* **2019**, *75*, 76–89. [CrossRef]
31. Bao, Y.; Tang, L.; Srinivasan, S.; Schnable, P.S. Field-based architectural traits characterisation of maize plant using time-of-flight 3D imaging. *Biosyst. Eng.* **2019**, *178*, 86–101. [CrossRef]

32. Honkavaara, E.; Saari, H.; Kaivosoja, J.; Pölönen, I.; Hakala, T.; Litkey, P.; Mäkynen, J.; Pesonen, L. Processing and assessment of spectrometric, stereoscopic imagery collected using a lightweight UAV spectral camera for precision agriculture. *Remote Sens.* **2013**, *5*, 5006–5039. [CrossRef]
33. Gonçalves, J.; Henriques, R. UAV photogrammetry for topographic monitoring of coastal areas. *ISPRS J. Photogramm. Remote Sens.* **2015**, *104*, 101–111. [CrossRef]
34. Xu, B.; Wang, N.; Chen, T.; Li, M. Empirical evaluation of rectified activations in convolutional network. *arXiv* **2015**, arXiv:1505.00853.
35. Luo, Q.; Zhang, J.; Hui, L. A Geocoding Method for Interferometric DEM in Difficult Mapping Areas. In Proceedings of the 31st of Asian Conference on Remote Sensing, Hanoi, Vietnam, 1–5 November 2010; pp. 1298–1304.
36. Luo, Y.; Qiu, X.; Dong, Q.; Fu, K. A Robust Stereo Positioning Solution for Multiview Spaceborne SAR Images Based on the Range–Doppler Model. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
37. Luo, Q.; Hu, M.; Zhao, Z.; Li, J.; Zeng, Z. Design and experiments of X-type artificial control targets for a UAV-LiDAR system. *Int. J. Remote Sens.* **2020**, *41*, 3307–3321. [CrossRef]
38. Xi, Y.; Luo, Q. A morphology-based method for building change detection using multi-temporal airborne LiDAR data. *Remote Sens. Lett.* **2018**, *9*, 131–139. [CrossRef]
39. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.