A Hierarchical Deformable Deep Neural Network and an Aerial Image Benchmark Dataset for Surface Multiview Stereo Reconstruction

Jiayi Li, *Senior Member, IEEE*, Xin Huang, *Senior Member, IEEE*, Yujin Feng, Zhen Ji, Shulei Zhang, and Dawei Wen

***Abstract*— Multiview stereo (MVS) aerial image depth estimation is a research frontier in the remote sensing field. Recent deep learning-based advances in close-range object reconstruction have suggested the great potential of this approach. Meanwhile, the deformation problem and the scale variation issue are also worthy of attention. These characteristics of aerial images limit the applicability of the current methods for aerial image depth estimation. Moreover, there are few available benchmark datasets for aerial image depth estimation. In this regard, this article describes a new benchmark dataset called the LuoJia-MVS dataset (http://irsip.whu.edu.cn/resources/resources_en_v2.php), as well as a new deep neural network known as the hierarchical deformable cascade MVS network (HDC-MVSNet). The LuoJia-MVS dataset contains 7972 five-view images with a spatial resolution of 10 cm, pixel-wise depths, and precise camera parameters, and was generated from an accurate digital surface model (DSM) built from thousands of stereo aerial images. In the HDC-MVSNet network, a new full-scale feature pyramid extraction module, a hierarchical set of 3-D convolutional blocks, and "true 3-D" deformable 3-D convolutional layers are specifically designed by considering the aforementioned characteristics of aerial images. Overall and ablation experiments on the WHU and LuoJia-MVS datasets validated the superiority of HDC-MVSNet over the current state-of-the-art MVS depth estimation methods and confirmed that the newly built dataset can provide an effective benchmark.**

Manuscript received 27 June 2022; revised 24 September 2022 and 14 December 2022; accepted 3 January 2023. Date of publication 5 January 2023; date of current version 17 January 2023. This work was supported in part by the Special Fund of the Hubei Luojia Laboratory under Grant 220100031, in part by the National Natural Science Foundation of China under Grant 42071311 and Grant 41971295, in part by the Foundation for Innovative Research Groups of the Natural Science Foundation of Hubei Province under Grant 2020CFA003, in part by the Wuhan 2022 Dawning Project under Grant 2022010801020123, and in part by the Wuhan University Experiment Technology Project Funding under Grant WHU-2020-SYJS-0007. *(Corresponding author: Xin Huang.)*

Jiayi Li is with the Hubei Luojia Laboratory, Wuhan 430079, China, and also with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: zjjercia@whu.edu.cn).

Xin Huang is with the School of Remote Sensing and Information Engineering, Wuhan University, the Hubei Luojia Laboratory, and also with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: huang_whu@163.com).

Yujin Feng, Zhen Ji, and Shulei Zhang are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: 2020282130098@whu.edu.cn; jz07@whu.edu.cn; zhangshulei@whu.edu.cn).

Dawei Wen is with the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430070, China (e-mail: daweiwen_whu@163.com).

Digital Object Identifier 10.1109/TGRS.2023.3234694

***Index Terms*— Deep learning, depth map-based stereo reconstruction, digital surface model (DSM), multiview stereo (MVS) reconstruction.**

I. Introduction

A. Motivations

OVER the past decades, multiview stereo (MVS) aerial image depth estimation has been a hot research field [1], [2]. At present, large-scale and highly accurate 3-D reconstruction of the Earth's surface is dominated by commercial software, such as Smart3D,[1] SURE [3], and Pix4D.[2] However, subject to the utilized conventional dense matching methods [4], [5], [6], [7], the software is at risk of false matching when dealing with scenarios with perspective distortion [8], and the required post-processing comes with a large manual labor cost [28]. Benefiting from the recent success of deep learning, learning-based MVS methods can produce higher-quality results than the conventional dense matching-based methods for close-range reconstruction [10], [11]. Thus, it is worthwhile introducing the learning-based MVS technique into aerial image depth estimation [12] as it has the advantage of being able to alleviate the above shortcomings [13].

B. Related Works

Benchmarks play a fundamental role in developing and evaluating MVS algorithms [14], [15]. To the best of our knowledge, there are only a few open-access MVS benchmark datasets: Middlebury [9], DTU [16], Tanks and Temples [17], ETH3D [18], BlendedMVS [19], and WHU [13]. The first five datasets are made up of close-range multiview images, which differ significantly from aerial images in view angle and camera parameters, and cannot be used as benchmarks for aerial image depth estimation. A recent dataset called the WHU dataset [13], which comes with accurate camera parameters and complete ground-truth depth maps, is the only multiview aerial image dataset created for aerial image depth estimation. The WHU dataset is a synthetic aerial image dataset, which was sampled from a region with an area of about 6.7 × 2.2 km, containing dense high-rise buildings,

[1] http://www.bentley.com/en/products/brands/contextcapture

[2] http://www.pix4d.com/

sparse factories, mountains covered by forest, bare land, and rivers [13]. Considering the limitations coming from dataset scarcity and insufficient land-cover diversity, it is still important to publish new datasets to support research in this field.

In recent years, many learning-based MVS deep neural networks have been developed [20], which can be divided into two main categories: voxel-based methods [21] and depth map-based methods [10]. The former methods parameterize a regular voxel mesh and learn a 3-D cost volume to build a 3-D scene model in an end-to-end manner. However, due to the huge memory cost of 3-D volumes, it is difficult to balance resolution and accuracy, and the process is limited by the error caused by spatial discretization [11]. According to a recent survey [22], the depth map-based approaches are superior due to their comprehensive consideration of efficiency and accuracy, although they are inherently 2.5-D. Moreover, as the digital surface model (DSM) is also an important product for aerial image-based applications, depth map-based techniques are worthy of attention.

Inspired by the binocular stereo matching deep neural networks [2], [22], the MVS network (MVSNet) network proposed by Yao et al. [10] introduces differentiable homography warping to encode the camera parameters and construct the 3-D cost volume, based on the plane sweep algorithm [23]. In MVSNet, the depth map-based neural network is composed of three modules: 1) feature extraction; 2) homography warping; and 3) 3-D cost volume regularization and depth regression. R-MVSNet [24], which was also proposed by the authors of MVSNet, utilizes a convolutional gated recurrent network instead of a 3-D convolutional neural network (3-D CNN [25]) to sequentially regularize the cost maps. Thus, R-MVSNet can achieve higher-resolution reconstruction at the same memory cost. MVSNet-Cas [26] decomposes the single cost volume into a cascade of multiple stages and introduces a coarse-to-fine regularization framework. The cascade networks that can keep the fine contextual information, as well as achieve high-resolution reconstruction, have become popular frameworks [27], [28]. The follow-up methods based on this paradigm modify one or several modules to improve the resolution or reduce the memory cost. For example, by using lightweight operators to replace the original convolutional [29], [30] or recurrent layers [31], [32], or sparsely sampling the data to be processed [33] by using the inherent spatial coherence of the depth maps (i.e., the core idea of PatchMatch [34]). However, in order to share the weights of the subsequent regularization module, each cost volume with different channels is transformed with a uniform 3-D CNN layer, which loses the multifeature information of the high-level cost volumes.

Although studies on close-range object reconstruction have made certain progress, the differences between close-range and aerial images still pose a challenge to the applicability of the above neural networks. RED-Net [13], which was the first network designed for MVS aerial image depth estimation, introduces a recurrent encoder–decoder architecture, instead of the stacking of three gated recurrent units (GRUs) in R-MVSNet, and has outperformed the conventional state-of-the-art MVS aerial image depth estimation methods. In contrast to MVSNet and R-MVSNet, RED-Net downsamples the output depth four times and reconstructs the depth map with a full resolution. More recently, MS-REDNet [28], which is a cascade version of RED-Net, further utilizes a high-resolution encoder–decoder module (i.e., U-Net [35]) to exploit the full resolution of the extracted features. However, both RED-Net and MS-REDNet struggle with the high GPU memory problem, and cannot scale up well to realistic imagery with large sizes. As a result, there is still much room for improvement.

C. Contributions

In this context, the objective of this research was to address the task of aerial image depth estimation by dealing with the shortcomings described above. Specifically, the contributions of this article can be summarized as follows:

1) We built a new large-scale open-source MVS aerial image dataset named the LuoJia-MVS dataset (named after the address of our university). The LuoJia-MVS dataset contains 7972 multiview units. Each unit is composed of five red-green-blue (RGB) images with a spatial size of 768×384 and a spatial resolution of 10 cm. Each image is also equipped with a depth map of the same size and a set of precise camera parameters. The dataset includes various land-cover types, such as cultivated land, forest, and residential land. The dataset can thus supplement the existing benchmark datasets, allowing more diverse aerial image depth estimation evaluation.
2) We propose a hierarchical deformable cascade MVS network (HDC-MVSNet) for aerial image depth estimation. HDC-MVSNet simultaneously undertakes high-resolution multiscale feature extraction and hierarchical cost volume module construction to generate full-resolution depth with abundant contextual information. First, with regard to the land-cover objects at varying scales and with a coarse spatial resolution, while the feature pyramid network (FPN) [37] in MVSNet-CAS can only fuse spatial information of two adjacent scales, a full-scale feature pyramid extraction module is employed to incorporate low-level details with high-level semantics from the feature maps in different scales. Second, on the basis of the cascade structure proposed in MVSNet-CAS, a hierarchical set of 3-D convolutional blocks is further constructed to leverage the multifeature information of the multistage cost volumes. Third, a deformable 3-D convolutional block is further applied to replace the regular 3-D convolutional block in the cost volume regularization module, to extend the 3-D receptive field with a small-size convolution kernel and deal with the deformation of the objects in multiview aerial images.

The rest of this article is organized as follows. Section II introduces the LuoJia-MVS dataset in detail. Section III then provides an overall review of the related deep neural networks. The proposed HDC-MVSNet method is introduced in detail in Section IV. Section V provides the experimental comparison as well as a discussion of the proposed HDC-MVSNet method

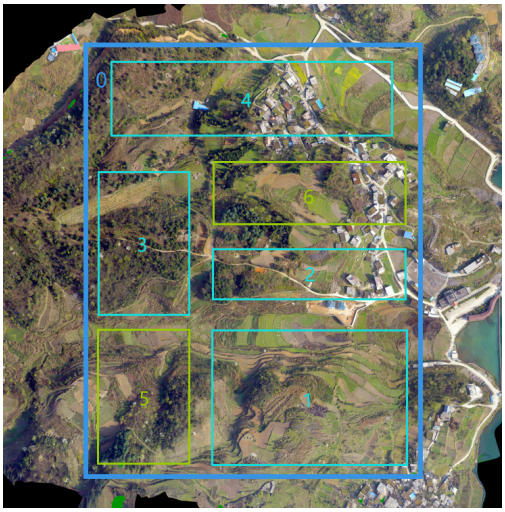


Fig. 1. LuoJia-MVS dataset. Area 0: the complete dataset consists of 7972 multiview units. Areas 1–4 were allocated for the training set, with 4320 multiview units, and areas 4 and 5 were selected for the test set, with 1360 multiview units.

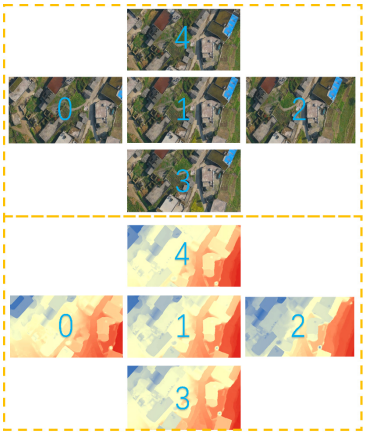


Fig. 2. Five-view unit with the size of 768 × 384. (Top) The RGB images. (Bottom) The corresponding depth maps. The three images tagged 0–2 make up the three-view set.

and the necessity for the LuoJia-MVS dataset. Finally, our conclusions are given in Section VI.

II. LuoJia-MVS Dataset

This section introduces the synthetic aerial image dataset called the LuoJia-MVS dataset for large-scale aerial image depth estimation, including the study area and data source, the construction process, and the data organization. As a five-view aerial image dataset, the LuoJia-MVS dataset follows the settings used in the construction of the WHU dataset [13].

A. Study Area and Data Source

Baiyun, Guiyang, Guizhou, China, was selected as the study area for the LuoJia-MVS dataset. As can be seen in Fig. 1, this is a hilly basin area dominated by mountains and hills, with an altitude of 1200–1600 m. The region contains a variety of land-cover types, including cultivated land, forest, urban areas, rural areas, industrial areas, mining areas, residential land, and unused land. Compared with the WHU dataset [13], the land-cover types are more diverse.

To construct the simulated multiview dataset, a 3-D DSM with OpenSceneGraph binary (OSGB) format mesh was built using a series of software tools, including Photoscan,[3] Smart3D, and Meshmixer,[4] from 1430 pairs of two-view aerial images. The size of each two-view image was 5304 × 7952 pixels and the spatial resolution was 5 cm. Manual editing was conducted to reduce the errors in the DSM.

B. Dataset Construction

In order to construct the five-view data units, we first simulated single-lens virtual aerial imagery with a given forward overlap and side overlap. The size, flight height, and spatial resolution of each virtual image were 960 × 480, 500 m above ground, and a 10-cm ground resolution, respectively, and both the forward overlap and side overlap were 90%. By setting the route, waypoint, and camera parameters, the metadata of each virtual image could be acquired. Each virtual image was then generated by the rendering engine, and the corresponding depth record was obtained.

The overlapping regions between all the single-lens virtual aerial images were then cropped and combined into units in the format of a five-view image set. We manually checked each unit and removed those with holes or deformation. In this way, a total of 7972 five-view units were generated. Each view consists of an RGB image with a size of 768 × 384 pixels and a resolution of 10 cm, along with its pixel-wise depth map, as shown in Fig. 2. The central image is the reference image, the images tagged 0 and 2 are in the forward direction, and the images tagged 3 and 4 in the side strips are the search images. Each image is near-nadir, with a specific intrinsic parameter matrix. In addition, a three-view subset with images tagged 0–2 was also collected.

Six representative sub-regions covering diverse landscapes were selected as the training and test sets, for which the spatial allocation is visually displayed in Fig. 1. Both the training and test set contain cultivated land, forest, urban areas, rural areas, industrial and mining areas, residential land, and unused land. A total of 4320 and 1360 multiview units form the training and test sets, respectively, in which the quantities are the same as for the WHU dataset. The ratio between these two sets is roughly 3:1. Thus, the LuoJia-MVS dataset complements the existing WHU dataset in terms of the land-cover types and dataset volume and further provides a new benchmark for MVS aerial image depth estimation. This dataset will be made publicly available for all research needs.

III. Fundamentals: The MVSNet-Cas Network

As with MVSNet, the basic idea of MVSNet-Cas is to construct a plane sweep volume [23] on the reference image and calculate the pixel-wise matching cost between the reference and the search images. Given a reference image and

[3]https://www.agisoft.com/
[4]https://www.autodesk.com/

its corresponding search images as input, MVSNet-Cas infers the depth of the reference image. The end-to-end MVSNet-Cas network consists of the following modules.

A. Feature Pyramid Extraction

A weight-sharing FPN [37] is built to extract the deep features of the input multiview images $\{\mathbf{O}_i\}_{i=1}^{N}$ (i.e., N views) for the subsequent dense matching, where $\mathbf{O}_i \in \mathbb{R}^{H\times W\times 3}$, and H, W, and 3 are the height, width, and number of bands (i.e., RGB) for the ith view image. In more detail, the multiscale feature maps with a spatial resolution of {1/16, 1/4, 1} of the original reference image size are separately employed to build three cost volumes with the corresponding resolutions. Here, we refer to the feature maps with the finest spatial resolution as those in the first stage, and the last stage deals with the feature maps with the coarsest resolution. In MVSNet-Cas, for the sake of efficiency, the outputs of this module are N F-channel feature maps $\{\mathbf{F}_i\}_{i=1}^{N}$, where $\mathbf{F}_i \in \mathbb{R}^{H\times W\times F}$.

B. Cascade Cost Volume Construction and Regularization

With the multiscale features from the FPN, the cascade cost volume utilizes the depth estimation with a coarse resolution to adaptively narrow the depth range. To maintain consistency with the feature extraction approach, the cost volume at the last stage is the coarsest, and the cost volume at the $(k-1)$th stage is built on the one at the kth stage. For the coarsest cost volume, based on the camera geometric projection, the plane sweep algorithm [23] is first utilized to warp the deep features of the search views into the coordinate system of the reference image

$$\mathbf{H}_i(d) = \mathbf{K_i}\cdot\mathbf{R}_i\cdot\left(\mathbf{I}-\frac{(\mathbf{t}_1-\mathbf{t}_i)\cdot n_1^T}{d}\right)\cdot\mathbf{R_1^T}\cdot\mathbf{K_1^{-1}},\quad i\in\{2,\ldots,N\} \tag{1}$$

where $\mathbf{H}_i(d) \in \mathbb{R}^{3\times 3}$ indicates the homography matrix between the feature maps of the search image i and the reference at depth d, and $\{\mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i\}$ and $\{\mathbf{K}_1, \mathbf{R}_1, \mathbf{t}_1\}$ are the intrinsic parameter matrix, rotation matrix, and translation vector of the search image i and reference camera, respectively. n_1 denotes the principal axis of the reference camera, and $\mathbf{I}$ is an identity matrix. For the sake of description, hereinafter, $\mathbf{H}_1$ is equal to $\mathbf{I}$, and $i \in \{1,\ldots,N\}$.

For the cascade cost volume construction, Fig. 3 illustrates the range reduction from stage 3 to 2. R_k and I_k respectively denote the depth range and hypothesis interval between the two adjacent hypothesis depth planes of stage k, and $D_k = R_k/I_k$. For stage 3, R_3 is equal to the full range of the whole reference image. The hypothesis depth interval I_3 is applied to generate a coarse depth estimation (the green lines in Fig. 3), which is leveraged to narrow the following range: $R_k = R_{k+1}\cdot v_k$, where $v_k \in \{0, 1\}$ is a hyperparameter at the kth stage. In this way, the range is gradually narrowed, and the efficiency is improved. Similarly, at the kth stage, another hyperparameter $u_k \in \{0, 1\}$ is also applied to recover more detailed depth variations by setting $I_k = I_{k+1}\cdot u_k$.

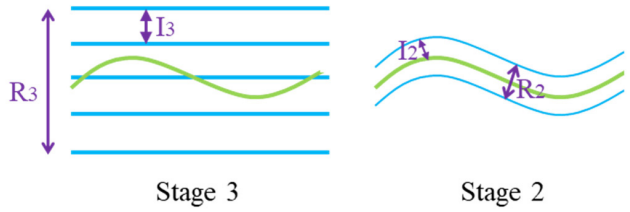


Fig. 3. Illustration of the adaptive hypothesis plane generation for stage 2. The blue lines are the hypothesis planes, and the green lines denote the predicted depth from stage 3. As shown in this figure, both the depth range and the interval can be narrowed, e.g., $I_2 = I_3 \times 0.5$ and $R_2 = R_3 \times 0.3$.

Then, as denoted by the green arrows in Fig. 4, except for the coarsest stage (i.e., $k \in \{1, 2\}$), the depth of the hypothesis planes at the kth stage is equal to the previous estimation at the $(k+1)$th stage (denoted as d_{k+1}) plus the depth residual of the current stage (denoted as Δ_k). $d_{k,\min}$ and $d_{k,\max}$ are, respectively, the minimum and maximum depth values of the depth range at the current stage R_k

$$d_{k,\min} = -0.5R_k,\quad d_{k,\max} = 0.5R_k. \tag{2}$$

For the jth interval, $\Delta_k^j = d_{k,\min} + I_k \times j$. Thus, the plane sweep algorithm-based cost volume construction [see (1)] in a multiscale manner can be expanded as

$$\mathbf{H}_i(d_{k+1}+\Delta_k) = \mathbf{K}_i\cdot\mathbf{R}_i\cdot\left(\mathbf{I}-\frac{(\mathbf{t}_1-\mathbf{t}_i)\cdot n_1^T}{d_{k+1}+\Delta_k}\right)\cdot\mathbf{R_1^T}\cdot\mathbf{K_1^{-1}}. \tag{3}$$

Accordingly, the cost volume at the kth stage (denoted as $\mathbf{C}_k$ with a size of $W_k \times H_k \times D_k \times F_k$) can be calculated as follows:

$$\mathbf{C}_k = \frac{\sum_{i=1}^{N}\left(\mathbf{V}_i^k - \overline{\mathbf{V}}_i^k\right)^2}{N} \tag{4}$$

where $\mathbf{V}_i^k$ and $\overline{\mathbf{V}_i^k}$ indicate the warped feature of the ith view image at the kth stage and the average of the multiview warped features, respectively.

A 3-D version of U-Net [10] is shared for each stage to transfer the F-channel $\mathbf{C}_k \in \mathbb{R}^{H\times W\times D}$ into the probability of the hypothesis depth plane $\mathbf{P}_k \in \mathbb{R}^{H\times W\times D}$. The 3-D U-Net model, which has the merit of aggregating neighboring information from a large receptive field, can reduce the amount of matching errors and keep the spatial smoothness. Finally, an ℓ_1-norm difference loss is incorporated after the softmax layer of the 3-D U-Net model. Finally, the weighted sum of the losses in all the stages constitutes the final loss of MVSNet-Cas, where each loss can be formulated as the ℓ_1-norm difference between the reference and the estimation.

C. Depth Estimation

If we suppose that $\mathbf{P}_k(d_k)$ is the estimated probability at depth d_k, then the depth estimation at stage $3\mathbf{dpt}_3 = \sum_{d_{3,\min}}^{d_{3,\max}} d_3 \times \mathbf{P}(d_3)$ and $\mathbf{dpt}_k = \sum_{\Delta_k^j}^{(\mathbf{dpt}_{k+1}+\Delta_k^j)\times\mathbf{P}}(\mathbf{dpt}_{k+1}+\Delta_k^j)$ when $k \in \{1, 2\}$. $\mathbf{Dpt}_k$ is the estimated depth at the kth stage.

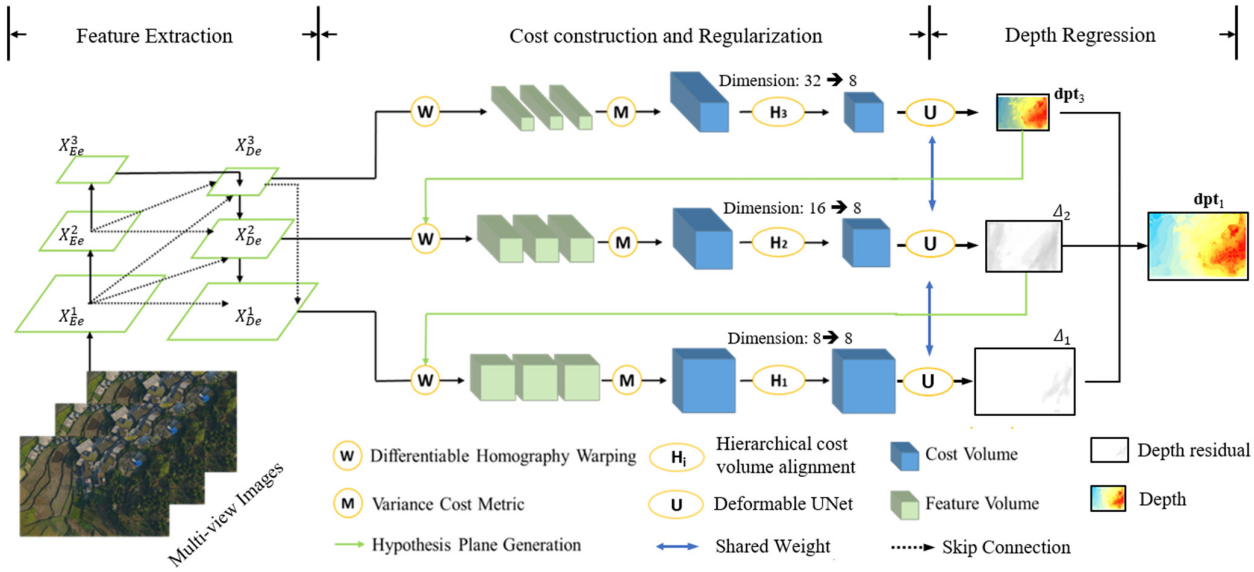


Fig. 4. Network architecture of the proposed HDC-MVSNet method. The proposed feature extraction module is shown in Figs. 5 and 6. The proposed detailed hierarchical cost volume alignment module is illustrated in Fig. 7. The detailed structure of the deformable 3-D U-Net is presented in Table I.

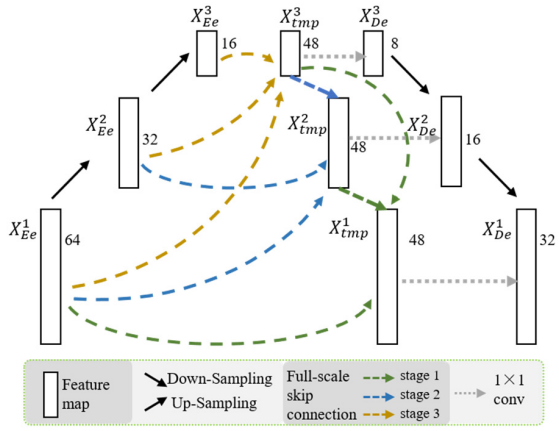


Fig. 5. Structure of the proposed feature extraction module. The number on the right side of each feature map indicates the feature channels.

IV. Proposed Network

In the following, we describe how we extended the cascade cost volume-based MVSNet and leveraged the multiscale and multifeature information of the multiview images for depth estimation (see Fig. 4). Specifically, the full-scale feature pyramid extraction module and the hierarchical deformable 3-D U-Net-based cost volume optimization are the major contributions.

A. Full-Scale Feature Pyramid Extraction Module

Inspired by the exchange unit aggregation of HRNet [38], the pyramidal adjacency of the FPN [37], and the full-scale connection of U-Net3+ [39], we propose a full-scale feature pyramid extraction module (see Fig. 5). This module is composed of an encoder, a temporary pyramid decoder, and a final semantic feature decoder. The encoder is equipped with three 2-D convolutional blocks, each of which consists of a convolutional layer sized 3×3 with stride 1, a max-pooling layer with stride 2, and a pool size of 2, and a rectified linear unit (ReLU) layer for non-linearity. Both the temporary and the final decoder have three stages, the features of which all have the same spatial size as the counterpart encoded features. The full-scale skip connection between the encoded features at multiple scales (denoted as X^i_{Ee} for scale i) and the temporary decoder feature (denoted as X^i_{tmp} for scale i) is employed to fuse the fine-grained and the coarser semantic information. In addition, the dense connection (formulated by a 1×1 convolution) between the temporary and final decoded feature (denoted as X^i_{DE} for scale i) is used to further aggregate the semantic information and reduce the channels of the final output. The numbers of channels for each feature are presented in Fig. 5. For the sake of comparison, the numbers of feature channels and the spatial size of both the encoder and final decoder are in line with those of the FPN in MVSNet-Cas [26]. Furthermore, the temporary feature at each scale has the same number of channels as in the decoder of U-Net3+ [39].

Fig. 6 illustrates the construction of the three representative skip connections, each of which incorporates the low-level encoded features with a finer spatial resolution into the decoded features. Specifically, the full-scale skip connection can capture the coarser semantic information and the fine-grained contextual variation at the same time, which is beneficial for land-cover types with varying scales. It should also be noted that, in contrast to the nested and dense connections in U-Net++ [40], the full-scale connection directly connects the encoded features with an equal or lower scale to the targeted decoded features. In this way, the full-scale skip connection is easier to train. The final decoded feature at each scale is then applied to the corresponding stage of the cascade cost volume optimization.

B. Hierarchical Deformable 3-D Convolution Module for Cascade Cost Volume Regularization

1) Cost Volume Feature Dimension Adaptation: As shown in Fig. 5, each decoder layer of the feature pyramid extraction

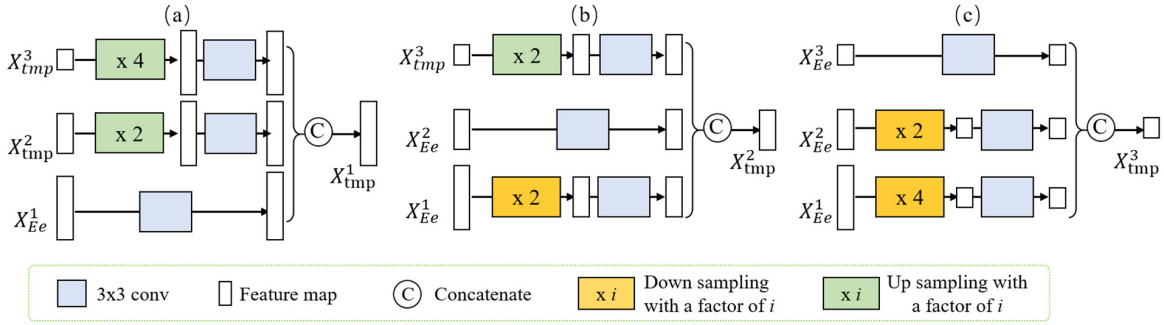


Fig. 6. Skip connections at (a) stage 1, (b) stage 2, and (c) stage 3. In this figure, the high-level features with a coarser spatial resolution are represented by rectangles with a shorter height.

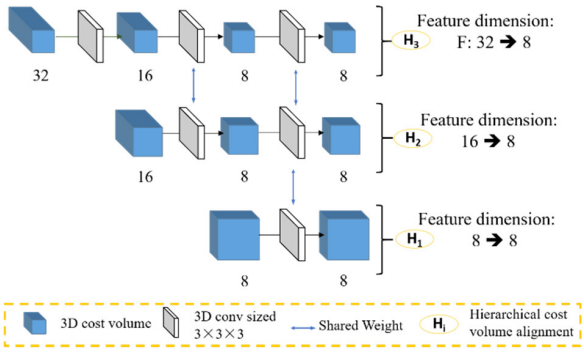


Fig. 7. Structure of the hierarchical cost volume alignment. The number below the blue cube indicates the feature dimension (i.e., the channels) of the cost volume.

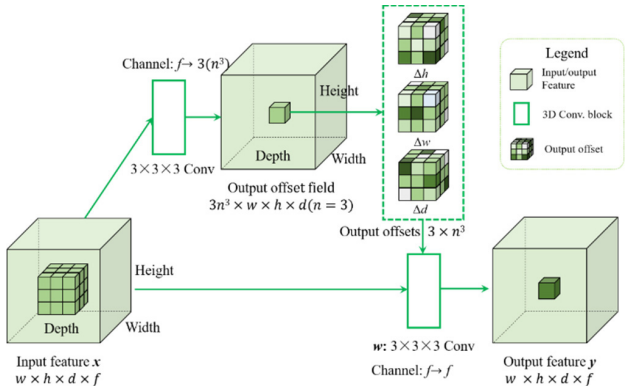


Fig. 8. Deformable "true 3-D" convolution. Please note that w, h, d, and f may take different values, depending on the stage. This convolution is only employed in the deformable 3-D U-Net module for cost volume regularization.

module (e.g., X^1_{De}–X^3_{De}) has different feature dimensions. MVSNet-Cas utilizes a single 3-D convolutional block with the kernel size of $3 \times 3 \times 3$ and different strides in each stage to transfer the hierarchical cost volumes (the dimensions for the three stages are 8, 16, and 32, respectively) to the same size (i.e., the dimension for each stage is 8). Thus, a considerable amount of cost volume information is lost.

To address this problem, a hierarchical set of 3-D convolutional blocks with a kernel size of $3 \times 3 \times 3$ is used to leverage the feature information and gradually unify the multistage cost volumes with the same feature dimension number. The detailed structure of this approach is shown in Fig. 7.

2) Deformable 3-D U-Net for Cost Volume Regularization: The four-scale 3-D version of U-Net, which applies a 3-D convolutional block as the basic unit, is a widely used module for cost volume regularization [10], [20], [26]. However, it should be noted that the modeling capability of the 3-D convolutional block is limited by the local perspective distortion in multiview Earth observations. In this research, we followed the architecture of the 3-D U-Net structure in MVSNet-Cas [26] and designed a deformable 3-D convolutional block instead of the standard 3-D convolutional block, to improve the cost volume regularization process.

To the best of our knowledge, the previous attempts at deformable 3-D convolution [41] still only learn the offsets on height and width. In this study, considering that geometric deformation can occur in each dimension of Euclidean space, a deformable "true 3-D" convolutional block with the offset at each dimension was designed. As shown in Fig. 8, the input cost volume feature with the size $w \times h \times d \times f$ is first fed into a plain 3-D convolution operator with the size of $3 \times 3 \times 3$ to generate offsets with the size of $3n^3 \times w \times h \times d$, with $n = 3$. These offsets have the same height-width-depth resolution as the input, while the channel dimension is set to $3n^3$ for the "true 3-D" deformable convolution. Each offset cube has n^3 channels, and its values indicate the deformation of the convolution sampling grid in width, height, and depth, i.e., Δw, Δh, and Δd. Then, with the guidance of these offsets, the plain 3-D convolution grid (i.e., the light green cubes in the input feature) can be moved to the deformable sampling grid (i.e., the dark green cubes in the input feature). In this way, the deformable 3-D convolution operation can be summarized as

$$\mathbf{y}(p_0) = \sum_{t=1}^{T} \boldsymbol{w}(p_t) \cdot \boldsymbol{x}(p_0 + p_t + \Delta p_t) \tag{5}$$

where p_0, p_t, and Δp_t respectively indicate location t in the output feature, the $3 \times 3 \times 3$ plain convolution sampling grid, and the offsets at location t. $\boldsymbol{y}$ and $\boldsymbol{x}$ are the output feature and input feature, respectively. In contrast to the previous 2-D works [41], [42], [43], trilinear interpolation is applied to calculate the feature $\boldsymbol{x}$.

A summary of the structure of the proposed deformable 3-D U-Net (i.e., the U module in Fig. 4), is provided in Table I.

TABLE I

STRUCTURE OF THE PROPOSED DEFORMABLE 3-D U-NET FOR COST VOLUME REGULARIZATION

Layer name	Output size	Output name	Operator
conv0x	8×H×W	init	[3×3×3 Pconv3d*, stride=1]×stage
conv1	16×(H/2)×(W/2)	down0	3×3×3 Pconv3d, stride=2
conv2			3×3×3 Pconv3d, stride=1
conv3	32×(H/4)×(W/4)	down1	3×3×3 Pconv3d, stride=2
conv4			3×3×3 Pconv3d, stride=1
conv5	64×(H/8)×(W/8)	down2	3×3×3 Pconv3d, stride=2
conv6			3×3×3 Pconv3d, stride=1
dconv3d	64×(H/8)×(W/8)	offset	3×3×3 Dconv3d**, stride=1
convd			3×3×3 Pconv3d, stride=1
\	32×(H/4)×(W/4)	up0	+ down2
conv7			3×3×3 Tconv3d***, stride=2
\	16×(H/2)×(W/2)	up1	+ down1
conv9			3×3×3 Tconv3d, stride=2
\	8×H×W	up2	+ down0
conv11			3×3×3 Tconv3d, stride=2
\	1×H×W	output	+ init
probx			[3×3×3 Pconv3d, stride=1]×4

*Pconv3d means plain 3D convolution, batch normalization, and a relu3D activation function
**Dconv3d means deformable 3D convolution, batch normalization, and a relu3D activation function
***Tconv3d means plain 3D transpose convolution, batch normalization, and a relu3D activation function

TABLE II

CHARACTERISTICS OF THE SEVEN DEPTH MAP-BASED MVS DEEP NEURAL NETWORKS

Method	GPU memory cost (MB)	Multi-stage technique	Neural layer in cost volume regularization
PatchmatchNet	5133	Cascade	3D conv
Fast-MVSNet	3942	Coarse to fine	3D sparse conv
MVSNet	6206	Single stage	3D conv
R-MVSNet	8959	Single stage	2D GRU
RED-Net	24339	Single stage	2D GRU
MVSNet-Cas	5490	Cascade	3D conv
HDC-MVSNet	7580	Cascade	3D conv and deformable 3D conv

Note: the GPU memory cost was estimated in a three-view scenario with each image sized 768 × 384.

V. EXPERIMENTS AND DISCUSSION

A. Datasets

Both the WHU dataset [13] and the LuoJia-MVS dataset built in this study were used in the experiments. Each dataset is made up of 4320 pairs of five-view images with a spatial resolution of 10 cm, each with a size of 768 × 384, and the ratio of the training set to test set is roughly 3:1. The major land-cover types of the WHU dataset are dense and tall buildings, sparse factories, mountains covered with forest, bare land, and rivers [13]. The land-cover types of the LuoJia-MVS dataset are cultivated land, forest, urban areas, rural areas, industrial and mining areas, residential land, and unused land.

B. Comparison Methods

To demonstrate the superiority of the proposed HDC-MVSNet method, the following six state-of-the-art depth-based MVS deep neural networks were employed in the comparison. The characteristics of these methods and the proposed method are listed in Table II. In this table, the first column lists the algorithm name, the second column lists the estimated GPU memory cost in a three-view scenario with each image sized 768 × 384, the third column lists the type of multistage technique, and the last column records the key neural modules used in each algorithm. The desirable depth estimation ability of MVSNet [10], MVSNet-Cas [26], R-MVSNet [24], and RED-Net [13] has been validated in recent studies [13], [28]. A detailed description of these networks can be found in Section III. Moreover, two recent high-efficiency multistage methods—PatchmatchNet [33] and Fast-MVSNet [29]—which can achieve appealing accuracies in close-range object reconstruction, were also considered.

PatchmatchNet is a cascaded deep neural network based on the learnable PatchMatch [34], which leverages random sampling to reduce the cost and spatial smoothness, to propagate the depth. The learnable PatchMatch is composed of a cost-matching step modeled by a 3-D CNN residual block with a kernel size of 1 × 1 × 1 and adaptive spatial cost aggregation modeled by a deformable 2-D convolutional block. Instead of the cascade structure, Fast-MVSNet [29] first reconstructs the depth of the coarse partial samples in the whole scene, then fills the holes under the guidance of the local smoothness prior, and finally refines the pixel-wise depth with an efficient Gauss–Newton layer, instead of a gradient descent layer.

C. Implementation and Accuracy Assessment

All the networks were run on a desktop computer using PyTorch 1.1.0 with an Intel Core i9-7980X CPU (2.60 GHz), 112-GB RAM, and an 11-GB GeForce RTX 2080Ti GPU. For all the methods, both three- and five-view reconstruction scenarios were tested to validate the robustness, while a batch size of one unit and the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ were adopted. The number of epochs was set to 30, with an initial learning rate of 0.001 for all the deep neural networks, and the learning rate was downscaled by a factor of 2 every two epochs after the number of epochs exceeded ten. In this study, only the depth map for the reference image was predicted.

For all the cascade structure networks, as suggested for the existing cascade-based methods, from the first to the third stage, the number of depth hypotheses D_1–D_3 was set to 48, 32, and 8, respectively; the depth intervals I_1–I_3 were set to 4, 2, and 1, respectively; the spatial resolution of the feature maps was set to {1/16, 1/4, 1} of the original reference image size; and the weight for each stage was set to be equal. The number of feature channels for the proposed network is presented in Figs. 6 and 8. In contrast, for a fair comparison, the depth hypotheses, the depth intervals, and the spatial resolution of the feature maps for the non-cascade structure networks were 192, 1, and the same as the original reference image size, respectively. The other parameters used in the comparison networks were in line with the original literature.

Three metrics are employed in this article to evaluate the performance as follows:

1) The mean absolute error (MAE) is used to assess the precision and is calculated as the average ℓ_1-norm difference between the true and estimated depths.
2) The <0.6 m (%) and <3-interval (%) indicators are used to assess the completeness, i.e., the percentage of pixels whose ℓ_1 error is less than 0.6 m and less than the three depth intervals. As the spatial resolution of both datasets in this study is 10 cm, the <3-interval here is equivalent to <0.3 m.

TABLE III

COMPARISON OF THE DIFFERENT DEPTH MAP-BASED MVS METHODS ON THE WHU DATASET

Number of views	Method	MAE (cm)	<3-interval (%)	<0.6 m (%)
Three-view	PatchmatchNet	17.3	94.8	96.5
	Fast-MVSNet	18.4	94.1	95.5
	MVSNet	19.0	94.3	95.0
	R-MVSNet	18.3	93.5	95.3
	RED-Net [13]*	11.2	**97.9**	**98.1**
	MVSNet-Cas	11.1	97.6	97.7
	HDC-MVSNet	**10.1**	97.8	97.9
Five-view	PatchmatchNet	16.0	95.0	96.9
	Fast-MVSNet	15.7	95.6	96.1
	MVSNet	16.0	95.5	95.8
	R-MVSNet	17.3	93.8	95.4
	RED-Net [13]*	10.4	97.9	**98.1**
	MVSNet-Cas	9.5	97.8	97.8
	HDC-MVSNet	**8.7**	**98.0**	**98.1**

RED-Net* indicates the accuracy from the original literature. RED-Net cannot be run with a single GeForce RTX 2080Ti GPU, due to its large memory requirement (i.e., more than 20 GB).

TABLE IV

COMPARISON OF THE DIFFERENT DEPTH MAP-BASED MVS METHODS ON THE LUOJIA-MVS DATASET

Number of views	Method	MAE (cm)	<3-interval (%)	<0.6 m (%)
Three-view	PatchmatchNet	25.5	87.2	92.7
	Fast-MVSNet	19.4	92.0	95.7
	MVSNet	17.2	92.4	96.1
	R-MVSNet	17.7	93.5	96.0
	MVSNet-Cas	10.3	97.1	98.4
	HDC-MVSNet	**8.9**	**97.8**	**98.7**
Five-view	PatchmatchNet	28.3	84.1	90.4
	Fast-MVSNet	35.7	74.9	84.6
	MVSNet	27.0	81.8	91.2
	R-MVSNet	25.9	86.7	92.3
	MVSNet-Cas	14.1	95.4	97.9
	HDC-MVSNet	**12.1**	**96.6**	**98.3**

D. Benchmark Performance

Tables III and IV list the reconstruction accuracies of the depth map-based MVS networks, with the best results in each test scenario highlighted in bold. It is demonstrated that the proposed HDC-MVSNet method achieves the best performance in both the three-view and five-view scenarios. In terms of the accuracy indicator, i.e., MAE, HDC-MVSNet outperforms all the other methods and obtains an improvement of at least 9% (i.e., $(11.1 - 10.1)/11.1 = 9\%$) on the WHU dataset and 13% on the LuoJia-MVS dataset, compared to the second-best MVSNet-Cas. In addition, the proposed HDC-MVSNet shows better completeness in terms of the <3-interval indicator. Although RED-Net shows desirable completeness in terms of the <0.6 m and <3-interval indicators, its large memory requirement (i.e., more than 20 GB in the three-view scenario with each image sized 768 × 384, as shown in Table II) indicates its limitation when applied to large-size aerial images.

MVSNet-Cas and the proposed HDC-MVSNet obtain much better accuracies than the other methods. Although the two most memory-efficient methods—PatchmatchNet and Fast-MVSNet—show competitive reconstruction accuracies when compared to MVSNet on the WHU dataset, they obtain the worst performances on the LuoJia-MVS dataset. It is noted that

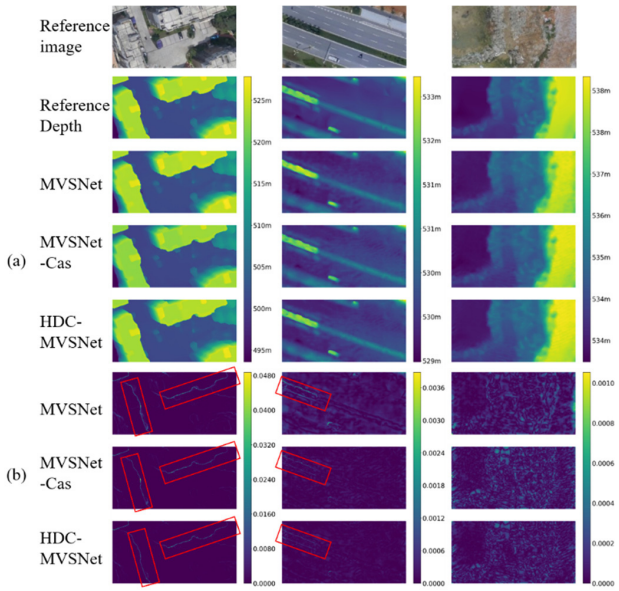


Fig. 9. Visualization comparison on three examples from the WHU dataset: (a) inferred depth maps and (b) corresponding relative depth residual maps for MVSNet, MVSNet-Cas, and the proposed HDC-MVSNet in the three-view scenario. Here, the relative residual is equal to the absolute difference between the reference depth and the inferred depth, divided by the reference depth. (Left) Tall building area, (middle) highway, and (right) unused land. Areas with large errors are marked with red rectangles.

both PatchmatchNet and Fast-MVSNet are multistage-based methods, while MVSNet is implemented in a single-stage manner, without any residual refinement. This phenomenon suggests that the spatial smoothness prior, which may be useful in close-range object reconstruction, does not handle the aerial image depth estimation task well, due to the coarser spatial resolution and the larger number of small above-ground objects. At the same time, the large gaps between R-MVSNet and MVSNet-Cas/HDC-MVSNet demonstrate the superiority of the cascade structure, which is an effective way of dealing with objects of varying scales.

Figs. 9 and 10 present several examples of the performance of HDC-MVSNet, MVSNet-Cas, and MVSNet on the WHU and LuoJia-MVS datasets, respectively, for a visual comparison. The typical land-cover types of the WHU dataset, i.e., tall building area, highway, and unused land, were selected, as shown in Fig. 9. The land-cover types that are rare in the WHU dataset, i.e., forest land, cropland, and rural residential area, were selected from the LuoJia-MVS dataset (see Fig. 10). As can be seen from these examples, all the networks show desirable performances for the homogenous regions of objects with a certain height, such as the tall buildings and the rural residential area. Objects with surface non-uniformities, such as forest land, have lower residuals than unused land. In addition, more residuals can be seen at the edges of the above-ground objects.

E. Ablation Study With the HDC-MVSNet Dataset

To verify the validity of the proposed full-scale feature pyramid extraction module and the hierarchical deformable 3-D U-Net, a series of ablation experiments were conducted on the two datasets (see Tables V and VI). In the following,

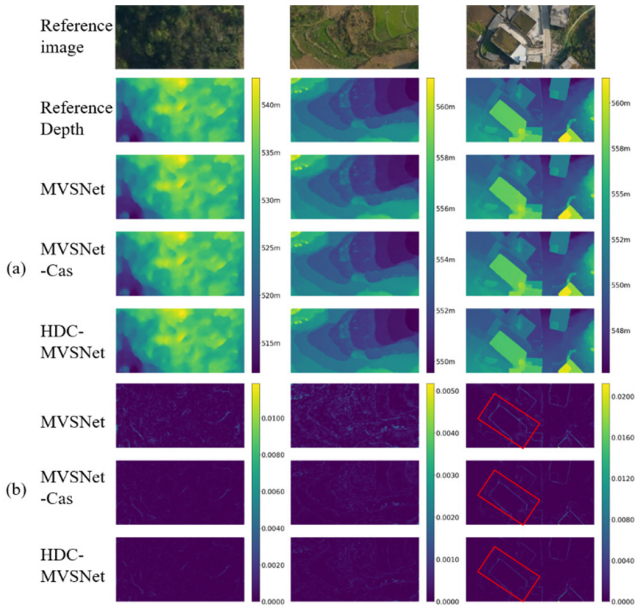


Fig. 10. Visualization comparison on three examples from the LuoJia-MVS dataset: (a) inferred depth maps and (b) corresponding relative depth residual maps for MVSNet, MVSNet-Cas, and the proposed HDC-MVSNet in the three-view scenario. Here, the relative residual is equal to the absolute difference between the reference depth and the inferred depth, divided by the reference depth. (Left) forest land, (middle) cropland, and (right) rural residential area. Areas with large errors are marked with red rectangles.

TABLE V

COMPARISON OF THE DIFFERENT DEPTH MAP-BASED MVS METHODS ON THE WHU DATASET IN THE THREE-VIEW SCENARIO

Cascade cost volume regularization module	Feature extraction module	MAE (cm)
3D U-Net (baseline)	*FPN (baseline)*	*11.1*
3D U-Net	Full-scale feature pyramid	10.6
Deformable 3D U-Net	FPN	10.8
Hierarchical deformable 3D U-Net	FPN	10.4
Hierarchical deformable 3D U-Net	FPN without cross-connections	10.5
Hierarchical deformable 3D U-Net	**Full-scale feature pyramid**	**10.1**

TABLE VI

COMPARISON OF THE DIFFERENT DEPTH MAP-BASED MVS METHODS ON THE LUOJIA-MVS DATASET IN THE THREE-VIEW SCENARIO

Cascade cost volume regularization module	Feature extraction module	MAE (cm)
3D U-Net (baseline)	*FPN (baseline)*	*10.3*
Hierarchical 3D U-Net*	FPN	9.6
Hierarchical 3D U-Net*	Full-scale feature pyramid	9.3
Deformable 3D U-Net	Full-scale feature pyramid	9.6
Hierarchical deformable 3D U-Net	FPN	9.4
Hierarchical deformable 3D U-Net	FPN without cross-connections	9.2
Hierarchical deformable 3D U-Net	**Full-scale feature pyramid**	**8.9**

*Hierarchical 3D U-Net indicates the combination of hierarchical cost volume alignment and 3D U-Net

"*baseline*" denotes the basic model, i.e., MVSNet-Cas, which uses the FPN as the feature extraction module and 3-D U-Net as the cascade cost regularization module. The last line of each table (in bold) denotes the proposed model, i.e., HDC-MVSNet, which uses the full-scale feature pyramid as the feature extractor module and hierarchical deformable 3-D U-Net for the cascade cost volume regularization. As can be seen in the last two rows in Tables V and VI, the cross-connections in the pyramid network can further aggregate the semantic information to achieve a superior MAE value.

TABLE VII

SENSITIVITY EXPERIMENTS WITH DIFFERENT DEPTH HYPOTHESIS NUMBERS AND DEPTH INTERVALS. THE STATISTICS WERE COLLECTED ON THE WHU AND LUOJIA-MVS DATASETS IN THE THREE-VIEW SCENARIO

Dataset	Stage num.	Depth num.	Depth interval	MAE (cm)	<3-interval (%)	<0.6 m (%)
WHU	1*	48	4	18.3	94.7	95.3
	2	48, 32	4, 2	11.8	97.4	97.5
	3	**48, 32, 8**	**4, 2, 1**	**10.1**	**97.8**	**97.9**
	4	48, 32, 8, 8	4, 2,1, 1	12.5	96.9	97.0
	4	48, 32, 8, 8	8, 4, 2, 1	12.6	96.9	97.1
LuoJia-MVS	1	48	4	21.2	88.8	94.8
	2	48, 32	4, 2	11.6	96.5	98.2
	3	**48, 32, 8**	**4, 2, 1**	**8.9**	**97.8**	**98.7**
	4	48, 32, 8, 8	4, 2, 1, 1	10.2	97.2	98.2
	4	48, 32, 8, 8	8, 4, 2, 1	10.0	97.5	98.4

1* indicates the baseline of this sensitivity experiment, which has no cascade cost volume.

As demonstrated in Tables V and VI, the incorporation of both the full-scale feature pyramid extraction module and the hierarchical deformable 3-D U-Net can improve the model performance on both datasets. More specifically, the full-scale feature pyramid extraction module and the hierarchical deformable 3-D U-Net obtain 4% and 6% MAE improvements on the WHU dataset, respectively, and the combination of these two modules results in a 13% MAE gain on the WHU dataset. For the LuoJia-MVS dataset, the incorporation of the hierarchical 3-D CNN layers results in a 6% MAE improvement, and a further 2.6% improvement is seen when replacing the plain 3-D U-Net with the deformable "true 3-D" layers. The further decomposition in the hierarchical deformable 3-D U-Net module suggests that the hierarchical 3-D CNN layers, which gradually aggregate the multifeature information, are beneficial. As shown in Table VI, the gain from 3-D U-Net to hierarchical 3-D U-Net is approximately equivalent to the combined increments of 3-D U-Net to deformable 3-D U-Net and FPN to full-scale feature pyramid. Moreover, when replacing the hierarchical 3-D U-Net with the deformable 3-D U-Net, there is even a drop in MAE (i.e., 9.31–9.6 cm).

F. Sensitivity Experiments for the Depth Hypothesis Numbers and Depth Intervals

The quantitative results with different stage numbers and depth intervals are summarized in Table VII. In this implementation, MVSNet with 48 depth hypotheses is employed as the baseline model, and the other versions replace its cost volume with the proposed hierarchical deformable cascade design, which is also composed of 48 depth hypotheses. In Table VII, the variants of the proposed HDC-MVSNet, which have different cascade stages, are denoted as "*i*," where *i* indicates the total stage number. Please note that the spatial resolutions of the different stages of each variant are the same as those of the baseline. It is demonstrated in Table VII that, as the number of stages increases, the accuracy indicators first increase greatly, then stabilize, and then slightly decrease when the number of stages exceeds three. According to the

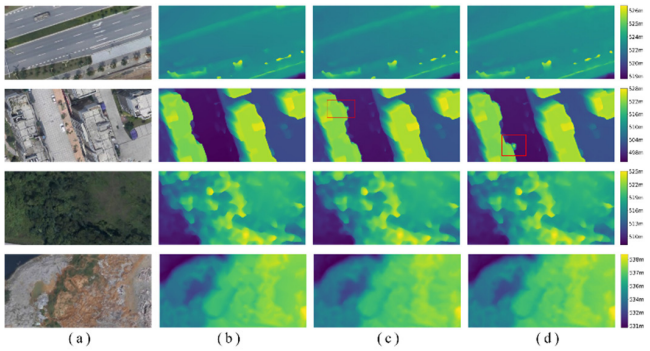


Fig. 11. HDC-MVSNet inference comparison using different views of the WHU dataset: (a) reference image, (b) reference depth, (c) depth estimation with the three-view dataset, and (d) depth estimation with the five-view dataset. Areas with large errors are marked with red boxes.

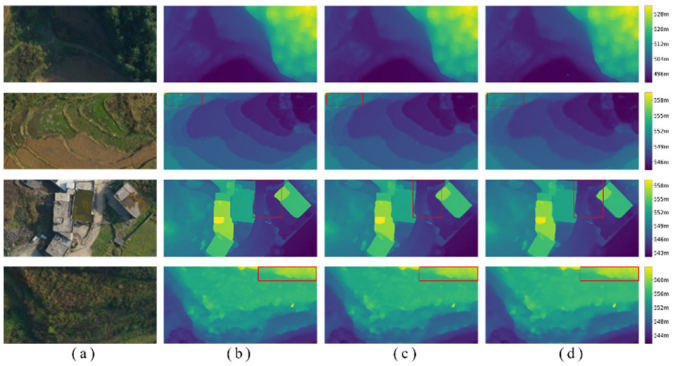


Fig. 12. HDC-MVSNet inference comparison using different views of the LuoJia-MVS dataset: (a) reference image, (b) reference depth, (c) depth estimation with the three-view dataset, and (d) depth estimation with the five-view dataset. Areas with large errors are marked with red boxes.

consistent experimental results obtained on both the WHU and LuoJia-MVS datasets, the suggested stage number is three.

G. Importance of the LUOJIA-MVS Dataset and Its Relationship With the WHU Dataset

The experiments on the LuoJia-MVS dataset proved its validity by means of a comprehensive comparison. However, we note that the reconstruction results for the five-view images are not as good as those for the three-view images, which is inconsistent with the performance obtained on the WHU dataset. From the perspective of the data sources, the DSM model used to simulate the LuoJia-MVS dataset was constructed based on 1430 pairs of two-view aerial images. Therefore, these results are likely due to the larger side overlap of the LuoJia-MVS dataset and the larger height-to-base ratio across the flying direction.

Moreover, comparisons using different views of both the LuoJia-MVS and WHU datasets are presented in Figs. 11 and 12. According to these two figures, it is apparent that:

1) In general, the aerial image depth estimation performances for these two datasets are satisfactory in both the three-view and five-view scenarios. It is also apparent that the main errors are located in the object edges [e.g., Figs. 9(b) and 10(b)] and the regions with large depth variations (e.g., the red boxes in the second row of Fig. 12), which is the case in both datasets.
2) In terms of the land cover, from the red boxes in the third row of Fig. 12, it can be seen that the reconstruction for the region suffering from shadow shows more errors. Therefore, it can be concluded that the LuoJia-MVS dataset is more challenging than the WHU dataset.

In summary, despite the complicated scenes and multiple land-cover types, relatively good depth estimation results can be obtained on the LuoJia-MVS dataset. The LuoJia-MVS dataset also contains several land-cover types that are not found in the WHU dataset. Moreover, there is also an urgent need to assess the reconstruction ability of complex regions, such as rural residential areas. For these situations, the LuoJia-MVS dataset can provide an effective benchmark.

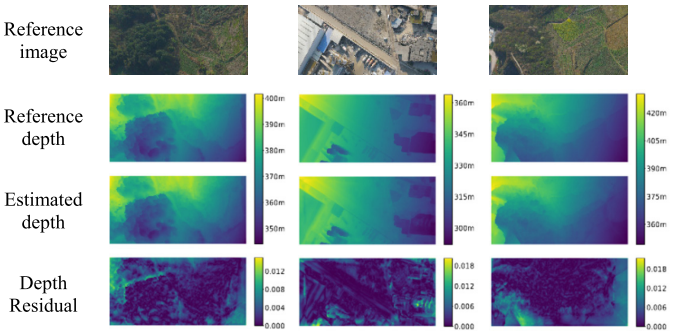


Fig. 13. Visualization comparison on three examples of real aerial images.

TABLE VIII
PARAMETERS OF THE REAL DATASET AND THE SYNTHETIC DATASET

Camera	Focal length	Forward overlap	Flying height	Near-nadir views?	Ground sampling distance
DSC-RX1RM2	35 mm	90%	1645.66 m	Yes	21 cm

H. Results Obtained on Real Aerial Imagery

In order to validate the effectiveness of HDC-MVSNet on real aerial imagery, three representative land-cover types were considered, i.e., forest land, factory, and rural residential area (see Fig. 13). To ensure sample independence, these images were also selected from the 1430 pairs of two-view aerial images, but were outside the spatial coverage of the LuoJia-MVS dataset. The parameters of these real images are listed in Table VIII. For each selected aerial image, Smart3D was used to correct the camera distortion, and Photoscan was used to estimate the camera parameters. The estimated depth of the real aerial images was then directly inferred by the HDC-MVSNet network trained on the proposed LuoJia-MVS simulated dataset, and the reference depth was estimated from the 3-D DSM mentioned in Section II-A.

In the three-view scenario, three real aerial images with certain forward overlaps were manually selected. In order to show the results for a larger area, to give a better visual impression, we down-sampled each aerial image by a factor of 4 and then cropped it into sub-regions with the size of 384 $\times$ 768. The forward overlap of these sub-images was no

less than 90%. In this way, the size of the real images was the same as the size of the simulated images in the LuoJia-MVS dataset. As can be seen in Fig. 13, the proposed HDC-MVSNet method still performs well on real aerial images.

VI. Conclusion

In this article, a new depth map-based MVS deep neural network named HDC-MVSNet has been proposed for depth estimation with multiview aerial images. We have also described a new benchmark dataset—the LuoJia-MVS dataset (http://irsip.whu.edu.cn/resources/resources_en_v2.php)—which complements the current aerial image depth estimation datasets in terms of the land-cover types, landscapes, and dataset volume. The first contribution of the proposed HDC-MVSNet method is the full-scale feature pyramid extraction module used to incorporate low-level details with high-level semantics from feature maps at different scales, which is applicable for the objects in aerial images. The second improvement refers to the hierarchical set of 3-D convolutional blocks, which are used to take advantage of the multifeature information, as the current algorithms do not consider the feature channels. Finally, a deformable "true 3-D" convolutional block was developed to deal with the deformation of the above-ground objects in multiview images. The experimental results demonstrated that the proposed HDC-MVSNet method outperformed the other state-of-the-art MVS aerial image depth estimation methods on both the WHU and LuoJia-MVS datasets. In the future, 3-D semantic detection, which integrates semantic segmentation and aerial image depth estimation as a uniform task, will be included in our agenda. Moreover, with regard to the pleasing performance of the depth map-based networks for aerial image depth estimation, it will be of interest to further develop this technique and apply it to the satellite image-based 3-D reconstruction task.

References

[1] Z. Rao, M. He, Z. Zhu, Y. Dai, and R. He, "Bidirectional guided attention network for 3-D semantic detection of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6138–6153, Jul. 2021.

[2] M. Mahato, S. Gedam, J. Joglekar, and K. M. Buddhiraju, "Dense stereo matching based on multiobjective fitness function—A genetic algorithm optimization approach for stereo correspondence," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3341–3353, Jun. 2019.

[3] M. Rothermel, K. Wenzel, D. Fritsch, and N. Haala, "SURE: Photogrammetric surface reconstruction from imagery," in *Proc. LC3D Workshop*, Berlin, Germany, vol. 8, no. 2, 2012, pp. 1–9.

[4] E. Tola, C. Strecha, and P. Fua, "Efficient large-scale multi-view stereo for ultra high-resolution image sets," *Mach. Vis. Appl.*, vol. 23, no. 5, pp. 903–920, 2012.

[5] M. Bleyer, C. Rhemann, and C. Rother, "Patchmatch stereo-stereo matching with slanted support windows," in *Proc. BMVC*, vol. 11, 2011, pp. 1–11.

[6] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2007.

[7] X. Huang, D. Wen, J. Li, and R. Qin, "Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery," *Remote Sens. Environ.*, vol. 196, pp. 56–75, Jul. 2017, doi: https://doi.org/10.1016/j.rse.2017.05.001.

[8] R. A. Beyer, O. Alexandrov, and S. Mcmichael, "The Ames stereo pipeline: NASA's open source software for deriving and processing terrain data," *Earth Space Sci.*, vol. 5, no. 9, pp. 537–548, Sep. 2018.

[9] D. Scharstein et al., "High-resolution stereo datasets with subpixel-accurate ground truth," in *Pattern Recognition* (Lecture Notes in Computer Science), vol. 8753, X. Jiang, J. Hornegger, and R. Koch, Eds. Cham, Switzerland: Springer, 2014.

[10] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 767–783.

[11] M. Blaha, C. Vogel, A. Richard, J. D. Wegner, T. Pock, and K. Schindler, "Large-scale semantic 3D reconstruction: An adaptive multi-resolution model for multi-class volumetric labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3176–3184.

[12] R. Chen, S. Han, J. Xu, and H. Su, "Point-based multi-view stereo network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1538–1547.

[13] J. Liu and S. Ji, "A novel recurrent encoder–decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6050–6059.

[14] J. Li, X. Huang, and J. Gong, "Deep neural network for remote-sensing image interpretation: Status and perspectives," *Nat. Sci. Rev.*, vol. 6, no. 6, pp. 1082–1086, May 2019.

[15] Y. Xu et al., "Luojia-HSSR: A high spatial–spectral resolution remote sensing dataset for land-cover classification with a new 3D-HRNet," *Geo-Spatial Inf. Sci.*, vol. 2, pp. 1–13, May 2022.

[16] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanaes, "Large scale multi-view stereopsis evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 406–413.

[17] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, 2017.

[18] T. Schops et al., "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3260–3269.

[19] Y. Yao et al., "BlendedMVS: A large-scale dataset for generalized multi-view stereo networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 1790–1799.

[20] K. Luo, T. Guan, L. Ju, H. Huang, and Y. Luo, "P-MVSNet: Learning patch-wise matching confidence aggregation for multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10452–10461.

[21] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang, "SurfaceNet: An end-to-end 3D neural network for multiview stereopsis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2307–2315.

[22] H. Laga, L. V. Jospin, F. Boussaid, and M. Bennamoun, "A survey on deep learning techniques for stereo-based depth estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1738–1764, Apr. 2022.

[23] R. T. Collins, "A space-sweep approach to true multi-image matching," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 1996, pp. 358–363.

[24] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent MVSNet for high-resolution multi-view stereo depth inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5525–5534.

[25] X. Huang et al., "A multispectral and multiangle 3-D convolutional neural network for the classification of ZY-3 satellite images over urban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10266–10285, Dec. 2021.

[26] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2495–2504.

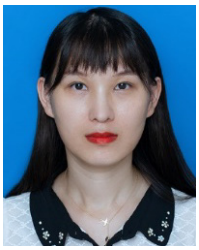
[27] J. Yang, W. Mao, J. M. Alvarez, and M. Liu, "Cost volume pyramid based depth inference for multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4877–4886.

[28] D. Yu, S. Ji, J. Liu, and S. Wei, "Automatic 3D building reconstruction from multi-view aerial images with deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 171, pp. 155–170, Jan. 2021.

[29] Z. Yu and S. Gao, "Fast-MVSNet: Sparse-to-dense multi-view stereo with learned propagation and Gauss–Newton refinement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1949–1958.

[30] Y. Wang, L. Wang, Z. Liang, J. Yang, W. An, and Y. Guo, "Occlusion-aware cost constructor for light field depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19809–19818.

[31] J. Yan et al., "Dense hybrid recurrent multi-view stereo net with dynamic consistency checking," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 674–689.
[32] R. Weilharter and F. Fraundorfer, "HighRes-MVSNet: A fast multi-view stereo network for dense 3D reconstruction from high-resolution images," *IEEE Access*, vol. 9, pp. 11306–11315, 2021.
[33] F. Wang, S. Galliani, C. Vogel, P. Speciale, and M. Pollefeys, "PatchmatchNet: Learned multi-view patchmatch stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14194–14203.
[34] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, Jul. 2009.
[35] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.*, 2015, pp. 234–241.
[36] Q. Zhu, C. Min, Z. Wei, Y. Chen, and G. Wang, "Deep learning for multi-view stereo via plane sweep: A survey," 2021, *arXiv:2106.15328*.
[37] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
[38] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696.
[39] H. Huang et al., "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Conf. Acoust., Speech Signal Process.*, May 2020, pp. 1055–1059.
[40] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.
[41] X. Ying, L. Wang, Y. Wang, W. Sheng, W. An, and Y. Guo, "Deformable 3D convolution for video super-resolution," *IEEE Signal Process. Lett.*, vol. 27, pp. 1500–1504, 2020.
[42] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Conf. Comput. Vis.*, Oct. 2017, pp. 764–773.
[43] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9308–9316.



Jiayi Li (Senior Member, IEEE) received the B.S. degree from Central South University, Changsha, China, in 2011, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2016.

She is currently an Associate Professor with the Hubei Luojia Laboratory, School of Remote Sensing and Information Engineering, Wuhan University. She has authored more than 60 peer-reviewed articles (Science Citation Index (SCI) articles) in international journals. Her research interests include hyperspectral imagery, sparse representation, computation vision and pattern recognition, and remote sensing images.

Dr. Li is a Reviewer of more than 30 international journals, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), IEEE TRANSACTIONS ON CYBERNETICS (TYCB), Remote Sensing of Environment (RSE), and *ISPRS-Journal of Photogrammetry and Remote Sensing (ISPRS-J)*. She is a Young Editorial Board Member of *Geo-Spatial Information Science* (GSIS), and a Guest Editor of the *Remote Sensing* (an open access journal from MDPI) and *Sustainability* (an open access journal from MDPI).

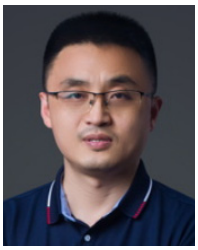
Xin Huang (Senior Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, in 2009.

He is currently a Full Professor with Hubei Luojia Laboratory, Wuhan University, where he teaches remote sensing and image interpretation. He is also the Head of the Institute of Remote Sensing Information Processing (IRSIP), School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 200 peer-reviewed articles (Science Citation Index (SCI) articles) in international journals. His research interests include remote sensing image processing methods and applications.

Prof. Huang was supported by the National Program for Support of Top-Notch Young Professionals in 2017, the China National Science Fund for Excellent Young Scholars in 2015, and the New Century Excellent Talents in University from the Ministry of Education of China in 2011. He was a recipient of the Boeing Award for the Best Paper in Image Analysis and Interpretation from the American Society for Photogrammetry and Remote Sensing (ASPRS) in 2010, the John I. Davidson President's Award from ASPRS in 2018, and the National Excellent Doctoral Dissertation Award of China in 2012. In 2011, he was recognized by the IEEE Geoscience and Remote Sensing Society (GRSS) as the Best Reviewer of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. He was the Winner of the IEEE GRSS Data Fusion Contest in 2014 and 2021. He was the Lead Guest Editor of the Special Issue of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, the *Journal of Applied Remote Sensing*, *Photogrammetric Engineering and Remote Sensing*, and *Remote Sensing*. He was an Associate Editor of the *Photogrammetric Engineering and Remote Sensing* from 2016 to 2019, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS from 2014 to 2020, and the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2018 to 2022; and has been serving as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING since 2022. He has also been an Editorial Board Member of the *Remote Sensing of Environment* since 2019.

Yujin Feng received the B.S. degree from the China University of Mining and Technology, Xuzhou, China, in 2020, and the M.S. degree in resources and environment from Wuhan University, Wuhan, China, in 2022.

His research interests include photogrammetry, land cover classification, semantic segmentation, and deep learning.



Zhen Ji received the B.S. degree in electronic engineering from Anhui University, Hefei, Anhui, China, in 2000, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2007.

He is currently an Associate Professor with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include photogrammetry and computer vision, close-range target recognition and positioning, real-time processing of low-altitude unmanned aerial vehicle (UAV) images, development of close-range and aerial photogrammetric software systems, and cultural relics digitization and 3-D modeling.

Shulei Zhang received the B.S. degree in remote sensing from Chang'an University, Xi'an, China, in 2021. She is currently pursuing the M.S. degree in remote sensing with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.

Her research interests include multiview stereo, building information extraction, high-resolution image processing, and deep learning.

Dawei Wen received the B.S. degree in surveying and mapping from Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, in 2018.

She is currently a Lecturer with the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan. Her research interests include urban remote sensing, high-resolution image processing, and change detection.