


# UAVStereo: A Multiple Resolution Dataset for Stereo Matching in UAV Scenarios

Xiaoyi Zhang , Xuefeng Cao , Anzhu Yu , Wenshuai Yu , *Member, IEEE*, Zhenqi Li, and Yujun Quan 

**Abstract**—Stereo matching is a fundamental task in 3-D scene reconstruction. Recently, deep learning-based methods have proven effective on some benchmark datasets, such as KITTI and SceneFlow. Unmanned aerial vehicles (UAVs) are commonly used for surface observation, and the images captured are frequently used for detailed 3-D reconstruction because of their high resolution and low-altitude acquisition. Currently, mainstream supervised learning networks require a significant amount of training data with ground-truth labels to learn model parameters. However, owing to the scarcity of UAV stereo-matching datasets, learning-based stereo matching methods in UAV scenarios are not fully investigated yet. To facilitate further research, this study proposes a pipeline for generating accurate and dense disparity maps using detailed meshes reconstructed based on UAV images and LiDAR point clouds. Through the proposed pipeline, we constructed a multiresolution UAV scenario dataset called UAVStereo, with over 34 000 stereo image pairs covering three typical scenes. To the best of our knowledge, UAVStereo is the first stereo matching dataset for UAV low-altitude scenarios. The dataset includes synthetic and real stereo pairs to enable generalization from the synthetic domain to the real domain. Furthermore, our UAVStereo dataset provides multiresolution and multiscene image pairs to accommodate various sensors and environments. In this article, we evaluate traditional and state-of-the-art deep learning methods, highlighting their limitations in addressing challenges in UAV scenarios and offering suggestions for future research.

**Index Terms**—Deep learning, disparity maps, stereo matching dataset, unmanned aerial vehicle (UAV).

## I. INTRODUCTION

ONE of the most active research areas in photogrammetry and computer vision is the 3-D reconstruction of environments via dense matching, which can be performed in stereo (in two views) [1] or multiview stereo (MVS) [2]. Among image-based approaches, stereo matching [3], in which expected correspondences are on epipolar lines, is arguably the most popular and intensively researched technique. Significant progress

has been made in this field in terms of the accuracy and cross-domain performance. Through stereo benchmarks [4], [5], [6], [7] researchers have achieved high accuracies on benchmarks for driving scenarios and indoor environments. Furthermore, some aerial stereo datasets have enabled deep learning to succeed in processing aerial stereo images [8], [9], [10]. However, the lack of large-scale datasets hinders research in terms of cross-domain performances and the application of stereo matching algorithms to unmanned aerial vehicle (UAV) images.

UAVs are a low-cost alternative to classical aerial photogrammetry for large-scale topographic mappings and even detailed 3-D recordings of ground information and are a valid complementary solution to terrestrial observations. With UAV images, current networks encounter three main challenges.

- 1) *Larger disparity search space*: The conversion between disparity and depth can be written as  $\text{disparity} = Bf/\text{Depth}$ , with baseline  $B$  and Depth generally in meters and the focal length in pixels  $f$ . The focal length in pixels  $f$  can be further expressed as  $W_{\text{px}}f_{\text{mm}}/W_{\text{CCD}}$ , where  $W_{\text{px}}$ ,  $f_{\text{mm}}$ ,  $W_{\text{CCD}}$  denote the image width in pixels, focal length in mm, and CCD width in mm, respectively. With the advancement of digital cameras, the resolution of acquired images is increasing, resulting in a larger disparity search space, which places increased computational performance requirements on the algorithm.
- 2) *Greater possibility of ill-areas*: UAVs are often used to capture information about ground surfaces, such as forests and grasslands, where features are difficult to match. In addition to the low-altitude acquisition characteristics, UAVs easily acquire images containing ill-areas, such as textureless and repetitive textures, which is extremely challenging and may be an inherently ill-posed problem in many cases.
- 3) *More varied disparity distribution*: Being lightweight and possessing adaptable characteristics, UAVs can collect images from various heights, where the disparity distribution significantly differs from driving and indoor scenarios and is significantly more varied. Most current algorithms are applied to datasets with the same disparity distribution, such as datasets on driving and indoor scenarios, which presents an additional challenge.

In addition to these properties of UAV images, the literature [5] has indicated that there is a significant difference in the performance of existing algorithms between the synthetic and real domains. To bridge this gap, synthetic image pairs have been used in advance for pretraining, and a small number of real

Manuscript received 19 December 2022; revised 5 February 2023; accepted 9 March 2023. Date of publication 15 March 2023; date of current version 29 March 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 42101458, 41801388, and 42130112. (Corresponding author: Xuefeng Cao.)

Xiaoyi Zhang, Xuefeng Cao, Anzhu Yu, Zhenqi Li, and Yujun Quan are with the PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China (e-mail: 1163635029@qq.com; cao\_xue\_feng@163.com; anzhu\_yu@126.com; Li13083833858@126.com; qj5312020@126.com).

Wenshuai Yu is with the The College of Civil and Transportation Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: ywsh@szu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3257489

Our dataset is available online at <https://github.com/rebecca0011/UAVStereo>. git

TABLE I  
COMPARISON OF AVAILABLE STEREO DATASETS

Dataset	Year	Scenario	Stereo number	Resolution	Disparity density
KITTI2012 [4]	2012	Driving	389	1226×370	sparse
KITTI2015 [30]	2015	Driving	400	1242×375	sparse
DrivingStereo [5]	2019	Driving	182188	1762×800	sparse
AppolloScape [31]	2019	Driving	19035	3384×2710	sparse
ETH3D [6]	2017	Indoor + Outdoor	47	940×490	dense
SatStereo [32]	2019	Aerial	72	1298×1286	dense
ISPRS2021 [8]	2021	Aerial	1092	1024×1024	sparse
WHUStereo [10]	2020	Aerial	10979	768×384	sparse
MPI Sintel [38]	2012	Synthetic animation	564	1024×436	dense
SceneFlow [21]	2016	Synthetic animation	26066	960×540	dense
Virtual KITTI [37]	2016	Synthetic driving	21260	1242×375	dense
<b>UAVStereo(Ours)</b>	<b>2022</b>	<b>Synthetic and real UAV</b>	<b>38781</b>	<b>Multiple Resolution</b>	<b>dense</b>

Notes: In bold is our UAVStereo dataset.

images are used for finetuning, which can significantly improve the pretrained models' capacity for real data. Evidently, neither current single synthetic datasets nor real datasets satisfy the requirements individually. Thus, synthetic and real data must be combined into one dataset.

To achieve this goal, we propose a UAV scenario dataset that contains both synthetic and real data. For the synthetic data, we propose a large-scale stereo dataset with sufficient variation, realism, and size to successfully train large networks in UAV scenarios. For the real data, the acquired images were provided after a four-step processing (including initial disparity map generation, epipolar image generation, epipolar disparity map generation, and postprocessing). In addition, the original resolution stereo pairs and corresponding disparity maps were used for the high-resolution network evaluation. The main contributions of this study are as follows.

- 1) We propose a pipeline (details are presented in Sections III-B and III-C) that can generate a dense disparity for both synthetic and real images from UAV-obtained images and point clouds.
- 2) We construct a new UAVStereo dataset, which consists of image pairs and dense disparity maps for three representative UAV scenes. To shorten the gap between the synthetic and real domains, we construct both synthetic and real data to increase the availability of both in real scenarios and decrease the quantity demand for real data. To adapt to the imaging characteristics and disparity distribution in UAV scenarios, we published multiresolution images and corresponding disparity maps.
- 3) We evaluate traditional and state-of-the-art deep techniques on our dataset. The results across different datasets and stereo methods demonstrate that our dataset is more suitable for UAV scenarios. Our dataset also presents challenges to current algorithms in terms of the resolution, disparity search range, and geospatial feature matching.

## II. RELATED WORK

### A. Stereo Matching Methods

For many years, most algorithms have solved the stereo-matching problem following a typical four-step pipeline [1]: matching cost computation, cost aggregation, disparity

computation, and disparity refinement. Among the vast literature on traditional algorithms [11] [12], [13], [14], [15], semiglobal matching (SGM) [16] is the most popular and is a reference approach combining mutual information and dynamic programming optimization in several directions [17].

With the development of deep learning, the early research efforts focused on replacing the individual steps of a conventional pipeline with deep learning counterparts. For instance, 2-D convolutional neural networks (CNNs) have proven effective in feature extraction [18]. SGM-Net uses a CNN to provide learned penalties for the SGM [19].

Subsequently, end-to-end deep stereo networks rapidly became the main focus [20], [21], [22]. Inspired by the FCN used in semantic segmentation [23], [24], DispNet [21] adopts an encoder-decoder architecture to enable end-to-end disparity regression, with which the matching cost can be directly integrated into the encoder volumes. GC-Net [20] combines contextual information with 3-D convolutions over a certain cost volume. PSMNet [25] integrates global context information using spatial pyramid pooling and regularizes the cost volume using stacked multiple hourglass 3-D convolutional networks. To address the high memory consumption of high-resolution image matching, DeepPrunner [26] proposed pruning the 3-D cost volume using a differential patch match method. STTR [27] revisited the problem from a sequence-to-sequence correspondence perspective and replaced cost-volume construction with dense pixel matching using position information and attention.

Although these methods were developed by the computer vision community on indoor or driving datasets, numerous researchers introduced geospatial aerial images into stereo matching networks, and this proved to be effective [28], [29].

Compared with traditional methods, learning-based stereo matching networks have shown excellent feature matching capabilities in many scenarios and have been applied in aerial image stereo matching because of their superior cross-domain generation capabilities. However, the capability of the network in UAV imagery has not been validated, owing to a lack of data.

### B. Stereo Benchmarks

The growing availability of datasets plays a crucial role in the rapid development of stereo matching. Table I lists some datasets

for stereo matching proposed by the aerial photogrammetry and computer vision communities. Some involve driving scenes: the KITTI datasets in two versions, KITTI 2012 [4] and KITTI 2015 [30]; and the large-scale stereo DrivingStereo [5] and AppolloScape [31] datasets. Moreover, Middlebury 2014 [7], which involves framing indoor environments, and ETH3D [6], containing both indoor and outdoor scenes, are also popular and widely utilized. In aerial photogrammetry, the SatStereo [32] and ISPRS2021 [8] datasets are frequently used for dense matching evaluations. For aerial datasets, UrbanScene3D [33] for aerial path planning and 3-D reconstruction, DFC19 [34] for large scene semantic 3-D reconstruction, INSANE [35] for cross-environment localization, and [36] for multiview 3-D reconstruction are also frequently used datasets. Using advanced computer graphics, the SceneFlow [21], Virtual KITTI [37], and MPI Sintel [38] datasets synthesize dense disparity maps; however, a significant gap remains between the synthetic domain and the real world.

UAVs are commonly used for Earth observation because of their mobility, flexibility, and low cost. However, UAV image processing requires considerable time and computational memory, owing to its high resolution. We are attempting to reduce the processing time by incorporating a stereo-matching network. Therefore, large-scale UAV scenario datasets are required to train the network. In this study, we propose the UAVStereo dataset, which contains a large number of image pairs with a dense disparity to facilitate the training and testing of stereo-matching networks.

### III. UAVSTEREO DATASET

This section introduces the UAVStereo data production process. Section III-A describes the data acquisition system and the areas covered. Sections III-B and III-C present the data production pipelines for the synthetic and real data, respectively.

#### A. Data Acquisition

The commercial UAV DJI Matrice 300<sup>1</sup> is a widely used platform for Earth observation. We chose DJI Matrice 300 as the platform, which was equipped with a Zenmuse L1 LiDAR sensor<sup>2</sup> and Zenmuse P1 full-frame imaging sensor<sup>3</sup> for obtaining point clouds and images, respectively, in the designated areas. The Zenmuse P1 imaging sensor has a  $35.9 \times 24$  mm full-frame sensor with a pixel size of  $4.4 \mu\text{m}$ , allowing the capture of high-quality photographs with a resolution of  $8192 \times 5460$  px. The Zenmuse L1 integrates a Livox LiDAR module and camera, allowing it to capture the details of complex structures and generate true-color point cloud models. The horizontal and vertical accuracies of the L1 radar were 10 and 5 cm, respectively. The maximum range of the DJI L1 is 190 m at 10%, 100 klx and 450 m at 80%, 0 klx.

The point clouds acquired by the Zenmuse L1 LiDAR were first converted into the standard las format using DJI Terra.<sup>4</sup>

Then, 3-D digital surface models in the OBJ format were reconstructed using Daspatial GET3D Cluster<sup>5</sup> from a substantial number of images and point clouds. The obtained images are first processed by feature points extraction, feature points matching, and aerial triangulation. We then evaluated the aerial triangulation result using the following metrics: the camera position reprojection error and the connection point reprojection error. To make the surface model more accurate to the actual scene, we manually deleted the images with excessive errors and their corresponding connection points. Finally, point clouds were added to connection points, to jointly generate the accurate textured model.

Three different scenarios were included in the UAVStereo dataset: residential land, forest, and mining areas. As shown in Fig. 1, the residential area contains dense and regular tall buildings, flat roadways, and other urban scene features, covering approximately  $700 \times 1200 \text{ m}^2$ . This area provides an urban scene with disparity saltation, such as buildings. The forest area contains high-coverage trees, several houses, and other field scene components, covering approximately  $1350 \times 1500 \text{ m}^2$ . This region has textureless and repeated-texture images, which presents difficulties for stereo-matching algorithms. The mining zone is composed of approximately  $700 \times 700 \text{ m}^2$  of agriculture, low structures, and bare ground, which contain a continuous variation of disparity. These three areas are representative areas for UAV Earth observations and can represent different disparity distributions.

#### B. Synthetic Dataset

The model trained on the synthesized data can provide a good initial pretrained model for application on real UAV images. Therefore, we generated multiresolution and multiscene UAV data to adapt to the application of different sensors and scenes. Similar to SceneFlow [21], we used the open-source 3-D creation suite Blender<sup>6</sup> to simulate the flight path of drones and render the results into tens of thousands of frames. As shown in Fig. 2, we rendered textured 3-D models into color images and corresponding ground-truth disparity maps, generating a synthetic dataset subset consisting of both low- and high-resolution subsets.

Given the intrinsic camera parameters [focal length  $f$ , principal point  $(x_0, y_0)$ ], render settings (image size  $W, H$ , and sensor size and format), exterior orientation [camera center  $(X_s, Y_s, Z_s)$ , and three rotational angles  $(\varphi, \omega, \kappa)$ , baseline  $B$ ], Blender could directly retrieve the depth of each pixel from the imported OBJ models. For stereo images, Blender's stereoscopy function allows users to observe models from a left and right stereoscopic perspective, so we generate stereo images using stereoscopy following the designed drone flight path. As for disparity, we adjusted the render settings and set the focal length, baseline and depth nodes and converted the depth to disparity through node calculation according to the formula  $\text{disparity} = Bf/\text{Depth}$  using the known configuration of the virtual stereo rig.

<sup>1</sup>[Online]. Available: <https://www.dji.com/au/matrice-300>

<sup>2</sup>[Online]. Available: <https://www.dji.com/au/zenmuse-l1>

<sup>3</sup>[Online]. Available: <https://www.dji.com/au/zenmuse-p1>

<sup>4</sup>[Online]. Available: <https://www.dji.com/au/dji-terra>

<sup>5</sup>[Online]. Available: <https://daspatial.com/>

<sup>6</sup>[Online]. Available: <https://www.blender.org/>

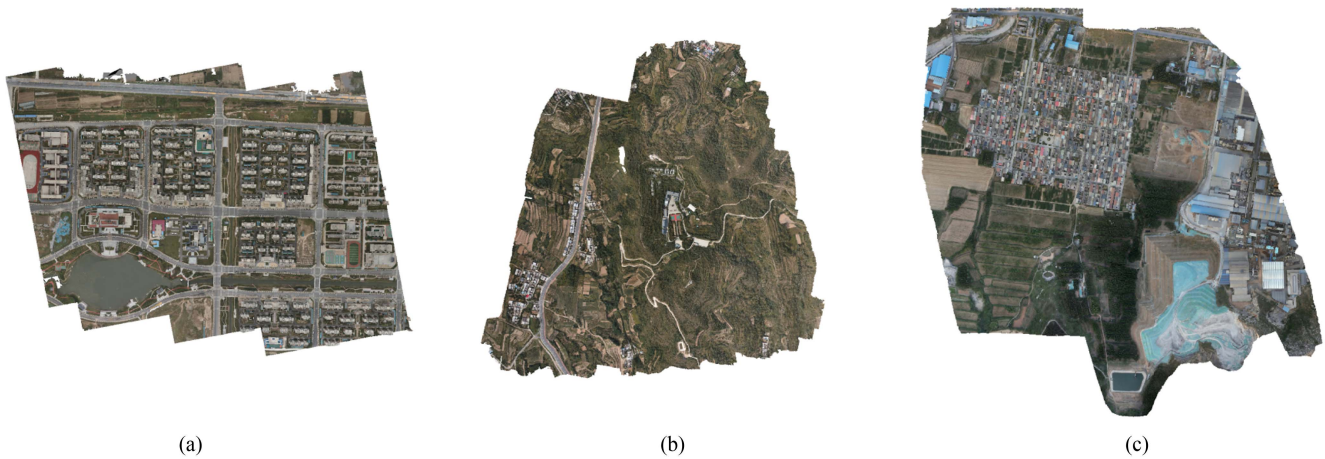


Fig. 1. Dataset acquisition areas. (a) Residential land. (b) Forest area. (c) Mining area.

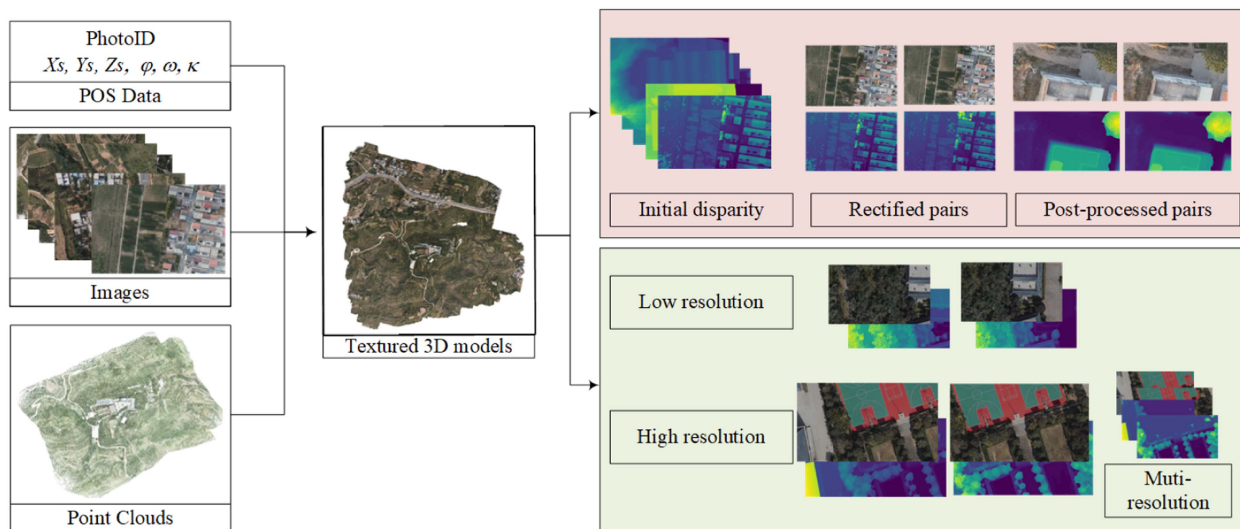


Fig. 2. Pipeline of dataset generation. The red box shows the real data generation process. The green box shows the synthetic data rendering process.

By adjusting the external orientation elements, the synthetic images and corresponding disparity maps were acquired at 100–300 m above the model with a high overlap. For all frames and views, we provided 8-bit RGB images and disparity maps in a portable float map (PFM) format for accurate and lossless disparity values. We rendered all image data using a virtual focal length of 35 mm on a 36-mm-wide simulated sensor. We released the high- and low-resolution subsets at  $960 \times 540$  px (same as SceneFlow) and  $8192 \times 5460$  px (same as Zenmuse P1). Simultaneously, we resized the high-resolution images to  $3840 \times 2160$  and  $1920 \times 1080$  px for the multi-resolution evaluation. The baseline was set within 1–15 m for the low-resolution subset owing to the image size limitation, and 15–35 m for the high-resolution subset. The image size, camera center, and baseline length of these two subsets were significantly different, which can provide multi-resolution images and test the robust performance of stereo-matching algorithms.

Fig. 3 presents the sample data in the synthetic subset with a baseline length of 15 m.

### C. Real Dataset

For real data, we used the images collected by P1, position and orientation system (POS) data containing image position and orientation, and georeferenced OBJ models. The collected data generated epipolar stereo image pairs and corresponding disparity maps in a four-step pipeline: initial disparity map generation, epipolar image generation, epipolar disparity map generation, and postprocessing.

After aligning the camera with the model using the position  $(X_s, Y_s, Z_s)$  and orientation  $(\varphi, \omega, \kappa)$  in the POS, an initial disparity map corresponding to the image can be rendered in the same manner as the synthetic data using Blender.

The second step of the processing pipeline is to create epipolar image pairs from adjacent images with sufficient overlap. Stereo

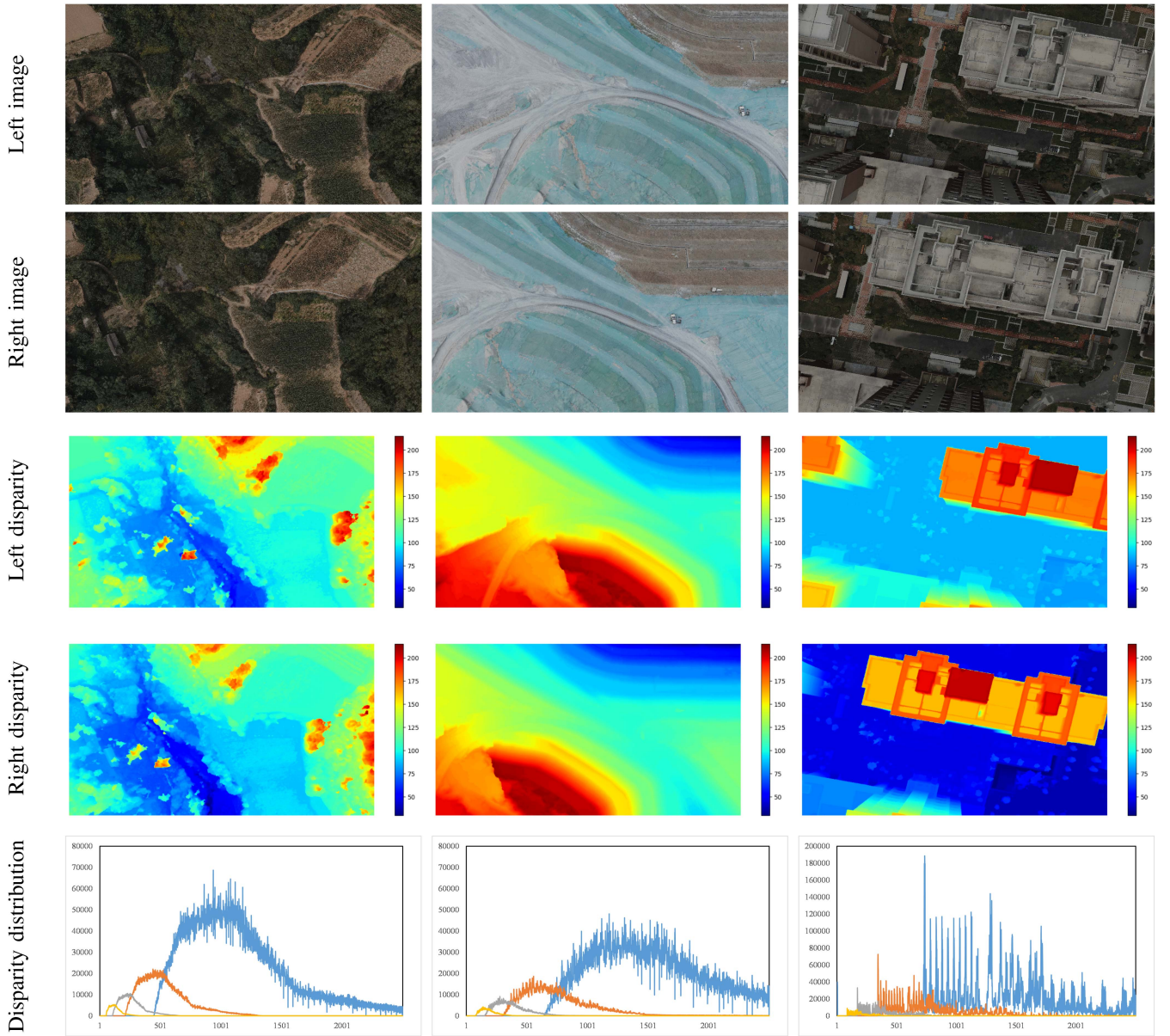


Fig. 3. UAVStereo Synthetic data. Left: Residential land. Center: Forest area. Right: Mining area. In disparity distribution histograms, the yellow, gray, orange, and blue lines, respectively, represent the disparity distribution of  $960 \times 540$ ,  $1920 \times 1080$ ,  $3840 \times 2160$ ,  $8192 \times 5460$  px disparity maps. The horizontal coordinate represents the disparity value (px) and the vertical coordinate represents pixel number.

rectification can be implemented using feature extraction and matching, fundamental matrix calculation, and interpolation resampling, which can implement off-the-shelf functions in the OpenCV library. The corresponding orientation parameters were generated to facilitate subsequent disparity map processing.

To retain the same transformation between pictures and disparity maps, we handled disparity maps by applying the orientation parameters from the previous step.

The well-chosen photos and corresponding disparity maps were then cropped to  $960 \times 540$  px, which is applicable to most networks.

Fig. 4 presents the resulting image pairs and corresponding disparity maps.

## IV. EXPERIMENTS

### A. Dataset Overview

The above process created a sizable stereo-matching dataset containing real and synthetic data using various scene photographs and LiDAR data collected by the UAV platform. We divided the training and testing sets at a ratio of approximately 8:2 for each scenario, splitting the total 38 781 stereo samples into 31 024 and 7757 for training and testing, respectively, as detailed in Table II.

In Table II, we list key parameters and additional details of the UAVStereo dataset.

The contributions of UAVStereo are summarized as follows.

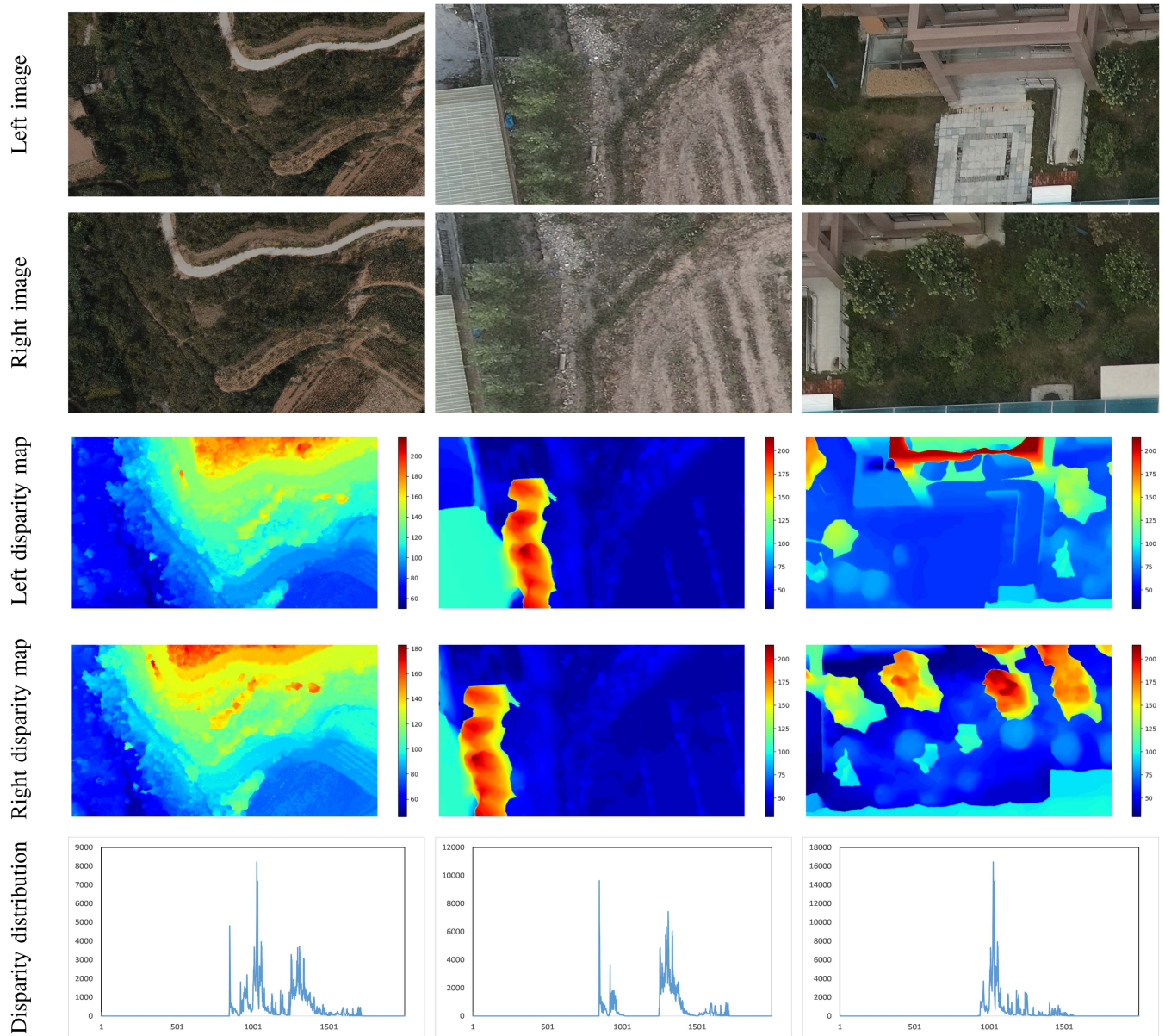


Fig. 4. UAVStereo Real data. Left: Residential land. Center: Forest area. Right: Mining area. In disparity distribution histograms, the horizontal coordinate represents the disparity value (px) and the vertical coordinate represents pixel number.

- 1) *First-ever UAV scenario stereo dataset*: Unlike autonomous driving, aerial, and indoor datasets, we propose a pipeline for generating images and disparity maps using UAV imagery and LiDAR point clouds in UAV scenarios.
- 2) *Large disparity range*: For changes in imaging sensor and exploring areas, our dataset contains multiple resolution images in representative scenes to adapt to the changes of payload sensors and environments.
- 3) *Containing both synthetic and real data*: Compared with the existing dataset containing only synthetic or real data, UAVStereo contains both synthetic and real data to bridge the gap between the real and synthetic domains.
- 4) *High diversity*: Our dataset provides various representative scenarios and multiple flight paths, making accounting

for most situations involving the top-down perspective of a drone possible.

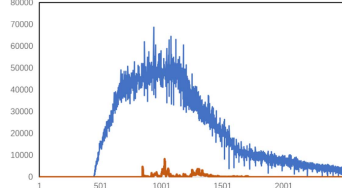
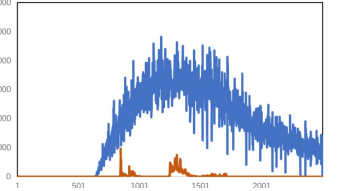
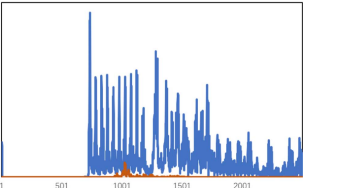
### B. Experiments Setup

After generating the epipolar images and corresponding ground-truth disparity maps, we evaluated several traditional and learning-based methods on our UAVStereo dataset.

In our experiments, we ran a series of deep-learning methods on UAVStereo to assess their accuracy, including PSM-Net [25], DSMNet [39], CFNet [40], RAFT-Stereo [41], and EAIStereo [42]. As a reference, we evaluated the popular SGM algorithm [16] as its fast variant.

We implemented the aforementioned deep neural network in PyTorch. All networks were trained end-to-end, using the images

TABLE II  
KEY PARAMETERS OF STEREO DATA IN UAVSTEREO

Scene type	Synthetic forest	Real forest	Synthetic mining	Real mining	Synthetic residential land	Real residential land
<b>Stereo number</b>	13320	1625	8640	3147	9000	3049
<b>Dataset split (Training / Testing)</b>	10656 / 2664	1300 / 325	6912 / 1728	2517 / 630	7200 / 1800	2439 / 610
<b>Baseline range</b>	15–35 <i>m</i>	22.8–35.6 <i>m</i>	15–35 <i>m</i>	26.7–33.7 <i>m</i>	15–35 <i>m</i>	14.6–19.2 <i>m</i>
<b>Resolution</b>	960 × 540 <i>px</i> 1920 × 1080 <i>px</i> 3840 × 2160 <i>px</i> 8192 × 5460 <i>px</i>	960 × 540 <i>px</i>	960 × 540 <i>px</i> 1920 × 1080 <i>px</i> 3840 × 2160 <i>px</i> 8192 × 5460 <i>px</i>	960 × 540 <i>px</i>	960 × 540 <i>px</i> 1920 × 1080 <i>px</i> 3840 × 2160 <i>px</i> 8192 × 5460 <i>px</i>	960 × 540 <i>px</i>
<b>Disparity distribution</b>						

Notes: In disparity distribution histograms, the orange and blue lines, respectively, represent the disparity distribution of real disparity maps (960 × 540 px) and synthetic disparity maps (8192 × 5460 px). The horizontal coordinate represents the disparity value (px) and the vertical coordinate represents pixel number.

as input, as well as disparity maps. We used the same learning rate, optimizer, and loss function as in the original network. Because deep network inferences are performed on an NVIDIA 3090 RTX GPU, we set the batch size to 1 and cropped the image to 256 × 512 px for all network training phases. The training process continued until the loss function no longer changed. The last epoch model was used for the evaluation.

The traditional SGM is available in OpenCV and is easy to implement using C++. In the SGM, we used a census to calculate the matching cost and aggregated the matching cost on eight paths. Postprocessing actions, such as consistency check, uniqueness constraint, and culling of small connected regions, were adopted for the completeness and consistency of the output.

To assess the accuracy of the stereo algorithms and networks, we used the endpoint error (EPE) and N-pixel error (N-PE) quantitative statistics as evaluation metrics. The EPE is the absolute mean of the difference between the estimated disparity map and ground truth, and N-PE is the percentage of pixels with an error larger than a threshold  $N$ :

$$\text{EPE} = \frac{1}{m} \sum_{i=1}^m |D_{\text{pred}} - D_{\text{gt}}|$$

$$\text{NPE} = \frac{\text{count}(|D_{\text{pred}} - D_{\text{gt}}| > N)}{m}$$

where  $m$  is the total number of pixels,  $D_{\text{pred}}$  is the output predicted disparity map, and  $D_{\text{gt}}$  is the ground-truth disparity map. As our ground-truth disparity maps were initially inferred at 960 × 540 px, we assumed three pixels as the lowest threshold. Then, given the much larger disparity in the real subset, we computed error rates up to 30- and 100-PE.

According to the dataset split in Table II, we trained deep neural networks on the training set until the loss function no

longer changed significantly, and calculated evaluation metrics on the testing set to assess the accuracy of networks. For the traditional SGM algorithm, we applied it to the testing set and calculated evaluation metrics.

### C. Evaluation on Synthetic Subset

In evaluating the synthetic subset, we trained the network on all training data instead of a single dataset to avoid overfitting on a particular subset. The pretrained models were verified under different scenarios. Table III compares the predicted disparity maps with its corresponding disparity ground truth.

All EPE results for UAVStereo were larger than those for other benchmarks, such as SceneFlow [21], KITTI [4], [30], and WHUStereo [10]. The primary reason for this is that disparity values are related to baseline values. UAV images have longer baselines than those of other datasets. According to the formula  $\text{disparity} = Bf/\text{Depth}$ , the longer the baseline is, the larger the disparity will be, which will lead to the increase of EPE error.

In evaluating the low resolution (upper portion of Table III), the results indicate that the traditional SGM method has considerable errors with UAV image pairs, whereas the end-to-end networks significantly improve the stereo-matching accuracy. When the SGM algorithm is applied to UAV images, the output disparity maps become incomplete and discontinuous, resulting in large error metrics. Comparing the performance of the SGM on Middlebury, we demonstrate that the SGM algorithm is unsuitable for processing UAV geographic images containing ill-areas, because this method uses a matching window of a limited size, which is incapable of obtaining or utilizing global information. Among the learning-based methods, PSMNet achieved the best results with EPE and 3-PE values of 3.443 px and 11.634%, respectively. The results demonstrate that our

TABLE III  
RESULTS ON THE UAVSTEREO SYNTHETIC SUBSET

Method	Evaluation Resolution	R		F		M		A	
		EPE (px)	3PE (%)	EPE (px)	3PE (%)	EPE (px)	3PE (%)	EPE (px)	3PE (%)
SGM [16]	960×540 px	102.035	92.577	135.767	96.125	69.484	92.756	102.428	93.819
PSMNet [25]	960×540 px	<b>4.688</b>	<b>15.701</b>	<b>4.421</b>	<b>15.057</b>	<b>4.084</b>	<b>15.031</b>	<b>3.443</b>	<b>11.634</b>
DSMNet [39]	960×540 px	9.003	53.434	6.861	44.306	5.317	35.918	4.482	24.423
CFNet [40]	960×540 px	14.995	48.621	5.509	39.131	10.019	36.383	7.371	42.554
RAFT-Stereo [41]	960×540 px	15.405	18.127	6.769	19.189	15.685	21.464	3.924	12.295
EAIStereo [42]	960×540 px	56.698	95.732	110.512	98.943	152.666	99.254	110.512	98.988
PSMNet [25]	1920×1080 px	<b>12.495</b>	28.090	13.280	<b>23.177</b>	14.346	29.795	10.274	21.558
DSMNet [39]	1920×1080 px	21.657	49.312	<b>7.295</b>	55.982	38.843	43.053	16.800	29.772
CFNet [40]	1920×1080 px	177.395	68.826	15.359	58.092	105.355	69.150	68.445	76.356
RAFT-Stereo [41]	1920×1080 px	30.806	<b>27.113</b>	13.530	28.766	<b>31.368</b>	<b>27.784</b>	<b>7.848</b>	<b>18.311</b>
PSMNet [25]	3840×2160 px	<b>28.485</b>	52.242	38.918	<b>26.221</b>	<b>24.691</b>	48.805	20.210	39.473
DSMNet [39]	3840×2160 px	36.550	59.702	<b>10.581</b>	62.935	69.968	45.078	29.923	38.377
CFNet [40]	3840×2160 px	217.121	69.205	30.718	73.103	179.721	78.928	137.193	87.522
RAFT-Stereo [41]	3840×2160 px	61.613	<b>42.941</b>	27.062	44.158	62.736	<b>40.351</b>	<b>15.697</b>	<b>29.806</b>

Notes: We trained model on 960×540 px and evaluated the model on multiple resolution ground-truth maps. Best scores in bold. R: Residential land testing subset, F: Forest areas testing subset, M: Mining areas testing subset, A: All synthetic testing set.

UAVStereo dataset can be applied to stereo-matching networks, with precision results comparable to those of blendedMVS [43] when converted into depth. Although the stereo-matching network has been rapidly developed since 2018, we determined that PSMNet performed best on low-resolution images for challenging data, such as low-altitude drone images, possibly because of the use of global information through spatial pyramid pooling and 3-D convolution strategies. Moreover, note that the loss function of the latest network, EAIStereo, converges to a large loss value, leading to large errors in the testing set. Consequently, EAIStereo may be unsuitable for UAV images.

Fig. 5 lists the representative disparity maps predicted by these algorithms. The disparity maps generated via SGM have a few invalid values, resulting in large error metrics. Learning-based models can infer complete and continuous disparity maps in challenging regions such as textureless ground. This result shows the superiority of deep-learning-based stereo matching on UAV images. Among the learning-based algorithms, PSMNet and RAFT-Stereo performed better on the dataset because accurate disparity maps could be obtained in all three scenarios.

The disparity search range increased with the image resolution. As shown in Figs. 3 and 4, the disparity search range of  $1920 \times 1080$  and  $3840 \times 2160$  px should be set to 768 and 960 px, causing an increase in computing memory usage. Because network inferences are performed on a single GPU, most algorithms can run only at  $960 \times 540$  px, owing to memory constraints. Consequently, their predicted disparities were upsampled with bilinear interpolation to compare with higher resolution ground-truth maps, with predicted disparities scaled by the upsampling factor itself. At the bottom portion of Table III, we list the evaluation results at  $1920 \times 1080$  and  $3840 \times 2160$  px. Because of the incompatibility of SGM and EAIStereo with drone imagery, we did not evaluate them at a larger resolution. We noted that all methods struggle to

TABLE IV  
RESULTS ON THE UAVSTEREO REAL SUBSET

	EPE (px)	30PE (%)	100PE (%)
Real	111.236	<b>71.867</b>	<b>48.288</b>
Finetuned	<b>101.386</b>	72.023	52.203

Notes: Three scenarios are used for the training set. Best scores in bold.

achieve good results at such a high resolution, with RAFT-Stereo achieving the best results. This result was expected because RAFT-Stereo achieved the top rank on Middlebury.

Moreover, in comparing the test results on the R, F, and M subsets, the error of PSMNet in the forest area is higher than those in the residential and mining areas, whereas the other methods are the opposite. This suggests that DSMNet, CFNet, and RAFT-Stereo are more suitable for the disparity estimation of repetitive textures, such as in forests.

#### D. Evaluation on Real Subset

To demonstrate the capacity of the real subset in UAVStereo, we selected the top-performing RAFT-stereo network from the previous evaluation, which can handle large disparity estimations. We conducted two experiments on the UAVStereo real subset: training the network directly with a real training set and finetuning the synthetic pretrained model using real data. In the latter experiments, we finetuned the pretraining model using only a quarter of the synthesized data. Considering the disparity range, we set the maximum disparity search range in the training stage of the real data to 1920 px. Owing to the large disparity value in the real scene, the EPE, 30-PE, and 100-PE were used to determine the inference error. In Table IV, we compared the metric errors of the pretrained model on the synthetic subset, the trained model on the real subset, and the finetuned model.



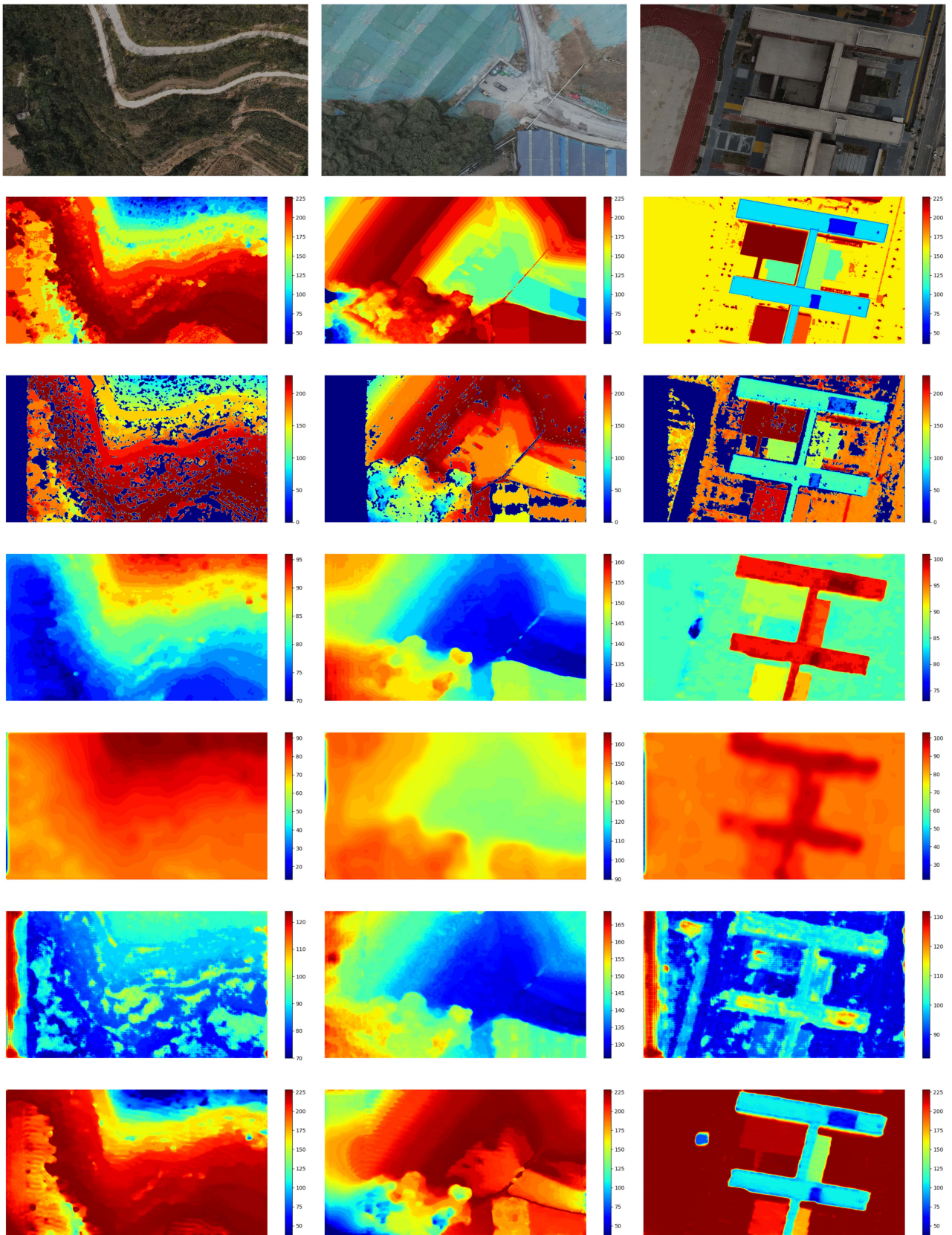


Fig. 5. Comparison of disparity maps estimated by different methods. From top to bottom: left image, ground-truth disparity map, SGM, PSMNet, DSMNet, CFNet, RAFT-Stereo.

Evidently, the real subset of UAVStereo can be used to train stereo matching models, although a large gap was observed in the results on the low-resolution synthetic datasets. This may be related to the large disparity in the search range. In addition, we determined that although utilizing less data, the finetuned results are comparable to the training results. This confirms our claims regarding the challenges in deep stereo networks, as well as the significance of our dataset.

### E. Discussion

UAVStereo is the first known UAV stereo matching dataset, filling a gap in the field. The proposed dataset introducing the use of learning-based techniques for UAV stereo matching. In addition, UAVStereo provides multiresolution data and multi-scene data, allowing it to better adapt to the resolution of UAV images and the variety of acquired scenes. What's more, we innovatively generated disparity maps corresponding to field images. The synthetic and real subsets provide the foundation data for the algorithm's generalization capabilities. We believe that UAVStereo can introduce learning-based stereo matching algorithms into UAV images, hence accelerating depth perception and 3-D reconstruction.

Several experiments indicated that UAVStereo can be used for stereo matching with both traditional and deep learning methods. By comparing the results on the low-resolution images (11.634% of 3-PE for the best result) with the results on KITTI benchmarks<sup>7</sup> (1.29% of 3-PE for the best result), we note how they still pose a significant challenge in geospatial UAV images owing to the large presence of ground objects. By comparing the results on the low-resolution images (11.634% of 3-PE, 3.443 px of EPE for the best result) with the results on high-resolution (18.311% of 3-PE, 7.295 px of EPE for the best result), we confirmed that the resolution is a challenge on our benchmark. In addition, it is observed that the results in the synthetic domain and the real domain are not of the same magnitude, and we believe that the current algorithm is not optimal for the processing of real UAV images, possibly as a result of factors such as the increase of the disparity searching range.

## V. CONCLUSION

The quantity and quality of datasets are critical aspects for the performance of stereo-matching algorithms. This study proposed a pipeline for generating image pairs and dense disparities for UAV scenes using images and point clouds. With the proposed pipeline, we constructed UAVStereo, a novel stereo dataset of UAV scenarios, containing synthetic and real image pairs and featuring a large disparity searching range and covering geospatial information, which is extremely challenging for existing learning-based networks. Although stereo datasets targeting autonomous driving, indoor, and aerial environments are available, UAVStereo is the first stereo-matching dataset of UAV scenarios, including a large number of image pairs and

dense labels, which should accelerate the 3-D reconstruction process.

UAVStereo dataset can represent, to an extent, the characteristics of UAV stereo-matching data with a large disparity search space, greater possibility of ill-areas, and more varied disparity distribution. Several experiments demonstrated that our dataset can be used for stereo matching of traditional and deep learning algorithms, and we determined that deep learning-based methods have significant advantages over traditional algorithms, showing great potential for development. Using a small amount of real data to finetune the pretrained model can help achieve accuracy comparable to that of real data, and this strategy can effectively reduce the amount of real data required.

Our experiments show that UAVStereo reveals some of the most intriguing challenges in deep stereo and opens up promising research directions. In particular, subsequent work fostered by UAVStereo may be devoted to the following:

- 1) investigating the ability of deep models to process large disparity search ranges;
- 2) enhancing the capability of processing geospatial data covering ground objects, which is significantly different from indoor driving scenes;
- 3) improving the generalization ability of a network between synthetic and real domains.

Although our UAVStereo dataset contains three typical scenes and numerous images, it needs to be supplemented by more surface features images. In addition, drone images with larger oblique angle should be generated to more comprehensively simulate the characteristics of drone images.

Above all, we hope that UAVStereo benefits future research on UAV scenario stereo matching and 3-D reconstruction.

## REFERENCES

- [1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1, pp. 7–42, 2002.
- [2] 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.*, 2014, pp. 406–413.
- [3] M. Poggi, F. Tosi, K. Batsos, P. Mordohai, and S. Mattoccia, "On the synergies between machine learning and binocular stereo for depth estimation from images: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5314–5334, Sep. 2021.
- [4] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.
- [5] G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, and B. Zhou, "DrivingStereo: A large-scale dataset for stereo matching in autonomous driving scenarios," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 899–908.
- [6] 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.*, 2017, pp. 3260–3269.
- [7] D. Scharstein et al., "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognit.*, 2014, pp. 31–42.
- [8] T. Wu, B. Vallet, M. Pierrot-Deseilligny, and E. Rupnik, "A new stereo dense matching benchmark dataset for deep learning," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 43, pp. 405–412, 2021.
- [9] M. Cournet et al., "Ground truth generation and disparity estimation for optical satellite imagery," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 43, pp. 127–134, 2020.

<sup>7</sup>[Online]. Available: [https://www.cvlibs.net/datasets/kitti/eval\\_stereo\\_flow.php?benchmark=stereo](https://www.cvlibs.net/datasets/kitti/eval_stereo_flow.php?benchmark=stereo)

- [10] 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.*, 2020, pp. 6050–6059.
- [11] L. De-Maetzu, S. Mattoccia, A. Villanueva, and R. Cabeza, "Linear stereo matching," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 1708–1715.
- [12] L. Di Stefano, M. Marchionni, and S. Mattoccia, "A fast area-based stereo matching algorithm," *Image Vis. Comput.*, vol. 22, no. 12, pp. 983–1005, 2004.
- [13] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 504–511, Feb. 2013.
- [14] Q. Yang, "A non-local cost aggregation method for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1402–1409.
- [15] K.-J. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 650–656, Apr. 2006.
- [16] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.
- [17] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 2, pp. 807–814.
- [18] J. Zbontar et al., "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2287–2318, 2016.
- [19] A. Seki and M. Pollefeys, "SGM-Nets: Semi-global matching with neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 231–240.
- [20] A. Kendall et al., "End-to-end learning of geometry and context for deep stereo regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 66–75.
- [21] N. Mayer et al., "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4040–4048.
- [22] J. Pang, W. Sun, J. S. Ren, C. Yang, and Q. Yan, "Cascade residual learning: A two-stage convolutional neural network for stereo matching," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 887–895.
- [23] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
- [24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
- [25] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5410–5418.
- [26] S. Duggal, S. Wang, W.-C. Ma, R. Hu, and R. Urtaun, "DeepPruner: Learning efficient stereo matching via differentiable PatchMatch," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4384–4393.
- [27] Z. Li et al., "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6197–6206.
- [28] P. Knöbelreiter, C. Vogel, and T. Pock, "Self-supervised learning for stereo reconstruction on aerial images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 4379–4382.
- [29] S. He, S. Li, S. Jiang, and W. Jiang, "HMSM-Net: Hierarchical multi-scale matching network for disparity estimation of high-resolution satellite stereo images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 188, pp. 314–330, 2022.
- [30] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3061–3070.
- [31] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The Apolloscape open dataset for autonomous driving and its application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2702–2719, Oct. 2020.
- [32] S. Patil, B. Comandur, T. Prakash, and A. C. Kak, "A new stereo benchmarking dataset for satellite images," 2019, *arXiv:1907.04404*.
- [33] L. Lin, Y. Liu, Y. Hu, X. Yan, K. Xie, and H. Huang, "Capturing, reconstructing, and simulating: The UrbanScene3D dataset," in *Proc. 17th Eur. Conf. Comput. Vis.*, 2022, pp. 93–109.
- [34] S. Kunwar et al., "Large-scale semantic 3-D reconstruction: Outcome of the 2019 IEEE GRSS data fusion contest-Part A," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 922–935, 2020.
- [35] C. Brommer et al., "INSANE: Cross-domain UAV data sets with increased number of sensors for developing advanced and novel estimators," 2022, *arXiv:2210.09114*.
- [36] M. Bosch, Z. Kurtz, S. Hagstrom, and M. Brown, "A multiple view stereo benchmark for satellite imagery," in *Proc. IEEE Appl. Imagery Pattern Recognit. Workshop*, 2016, pp. 1–9.
- [37] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4340–4349.
- [38] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 611–625.
- [39] F. Zhang, X. Qi, R. Yang, V. Prisacariu, B. Wah, and P. Torr, "Domain-invariant stereo matching networks," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 420–439.
- [40] Z. Shen, Y. Dai, and Z. Rao, "CFNet: Cascade and fused cost volume for robust stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13906–13915.
- [41] L. Lipson, Z. Teed, and J. Deng, "RAFT-stereo: Multilevel recurrent field transforms for stereo matching," in *Proc. Int. Conf. 3D Vis.*, 2021, pp. 218–227.
- [42] H. Zhao, H. Zhou, Y. Zhang, Y. Zhao, Y. Yang, and T. Ouyang, "EAI-stereo: Error aware iterative network for stereo matching," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 315–332.
- [43] Y. Yao et al., "BlendedMVS: A large-scale dataset for generalized multi-view stereo networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1790–1799.



**Xiaoyi Zhang** received the bachelor's degree in remote sensing science and technology from the School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo, China, in 2020. She is currently working toward the master's degree with the PLA Strategic Support Force Information Engineering University, Zhengzhou, China. Her main research focuses on 3-D Scene reconstruction with UAV Images.



**Xuefeng Cao** received the bachelor's, master's, and Ph.D. degrees in cartography and geographic information system from the Institute of Surveying and Mapping, PLA Strategic Support Force Information Engineering University, Zhengzhou, China, in 2006, 2009, and 2012, respectively. He is currently an Associate Professor with the PLA Strategic Support Force Information Engineering University. His research interests include signal processing and industry applications in Earth observation.



**Anzhu Yu** received the bachelor's degree in remote sensing science and technology and the master's degree in photogrammetry and remote sensing from the PLA Strategic Support Force Information Engineering University, Zhengzhou, China, in 2011 and 2014, respectively, and the Ph.D. degree in photogrammetry and remote sensing from the Institute of Surveying and Mapping, PLA Strategic Support Force Information Engineering University, in 2017. He is currently an Associate Professor with the PLA Strategic Support Force Information Engineering University. His research focuses on signal processing in Earth observation.



**Wenshuai Yu** (Member, IEEE) received the bachelor's degree in aerospace photogrammetry, and the master's and Ph.D. degrees in photogrammetry and remote sensing from the Institute of Survey and Mapping, Zhengzhou, China, in 2003 and 2006, respectively.

He is currently an Associate Researcher with the College of Civil and Transportation Engineering, Shenzhen University, Shenzhen, China, and a Principal Investigator with the Guangdong Laboratory of Artificial Intelligence and Digital Economy, Shenzhen. His research interests include UAV photogrammetry, intelligent 3-D spatial perception, SLAM, and 3-D reconstruction.



**Yujun Quan** received the bachelor's degree in surveying engineering from the North China University of Water Resources and Electric Power University, Zhengzhou, China, in 2021. She is currently working toward the master's degree with the PLA Strategic Support Force Information Engineering University, Zhengzhou.

Her main research focuses on building extraction based on aerial and satellite remote sensing images.



**Zhenqi Li** received the bachelor's degree in remote sensing science and technology from School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo, China, in 2020. He is currently working toward the master's degree with the PLA Strategic Support Force Information Engineering University, Zhengzhou, China.

His main research focuses on viewpoints and path-planning for UAV-based 3-D reconstruction.