Surface Depth Estimation From Multiview Stereo Satellite Images With Distribution Contrast Network

Ziyang Chen, *Student Member, IEEE*, Wenting Li, Zhongwei Cui, and Yongjun Zhang, *Member, IEEE*

***Abstract*—The calculation of surface depth based on multiview stereo (MVS) satellite imagery is of significant importance in fields such as military and surveying. The challenge in extracting depth information from satellite imagery lies in the fact that these images often exhibit similar colors, necessitating the development of algorithms that can integrate shape and texture information. Moreover, the application of classical convolutional neural network (CNN) MVS is limited by its inability to capture long-range terrain relationships, which presents a bottleneck in existing surface depth estimation algorithms. To address the above problems, we propose the Distribution Contrast Network for Surface Depth Estimation from Satellite MultiView Stereo Images (DC-SatMVS), a novel satellite MVS network. In order to learn short-range and long-range features, we designed separate CNN and ViT branches. To emphasize the importance of shape and texture, we propose the Distribution Contrast Loss mechanism. This mechanism supervises the model training based on the similarity between the predicted depth and the ground truth depth distribution. Experimental results demonstrate that our method achieves state-of-the-art (SOTA) performance. We produce a remarkable 18.14% reduction in root mean square error compared to the Sat-MVSF on the WHU-TLC dataset. To validate the generalization performance of our framework, we trained and tested it on the DTU dataset, a common MVS dataset, and achieve SOTA results in this dataset as well.**

***Index Terms*—Multiview stereo (MVS), satellite stereo reconstruction, surface depth estimation.**

I. Introduction

WITHIN computer vision and remote sensing, the field of surface depth estimation [1] from multiview [2] optical pictures is an important and rapidly developing field [3]. According to [4], existing techniques for estimating surface depth from satellite imagery [5] can be divided into two main classes: One based on manual matching approaches using commercial software, i.e., ArcGIS, Catalyst, or open-source solutions [6], and the other based on deep learning-based multiview stereo (MVS) methods [3], [4], [7]. Traditional manual matching approaches require the human imposition of prior conditions, making the cost of matching prohibitive for complex scenes. In contrast, deep learning-based solutions [7], [8], [9] have demonstrated superior results in both remote sensing and general image contexts, making these approaches more widely applicable to research.

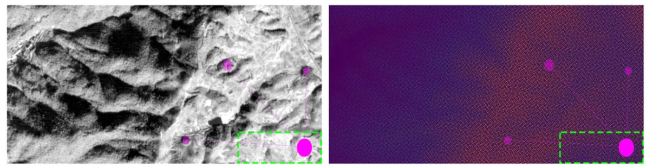

Fig. 1. **Motivations:** First, **The local geometry of satellite imagery is more similar**. Incorporating information from longer distances to discern the semantic patterns of the terrain proves to be a superior choice for depth estimation. The pink circles in this figure simulate the interaction of longrange sensory fields. Second, **Scenes with similar colors may have different depths**. As can be seen from the contents of the green box, satellite imagery has little color differentiation. Therefore, remote sensing MVS should pay more attention to features that are not related to color. The left half of Fig. 1 is the original image, and the right half is the ground truth.

As a representative work in satellite MVS, RED-Net [10] employs a recurrent convolutional neural network (CNN) [11] architecture and represents a pivotal contribution specifically designed for satellite MVS. CasMVSNet [7] decomposed the single-cost volume into a cascaded structure of multiple stages. It leveraged the depth map from the preceding stage to refine the depth range for each subsequent stage. UCS-Net [9] presented a MVS reconstruction method based on adaptive volume representation and uncertainty perception. The aforementioned algorithm has achieved promising results in diverse MVS scenarios [12]. Gao et al. [3] extended the applicability of CasMVSNet [7] and UCS-Net [9] to remote sensing imagery scenarios, demonstrating the effectiveness of them in satellite MVS applications. In order to better adapt to large-scale earth surface reconstruction, Gao et al. [3] proposed the Rational Polynomial Camera (RPC) Distortion Module to enhance existing satellite MVS methods. Furthermore, Gao et al. [4] introduced, Sat-MVSF, a more refined workflow [4] aimed at reducing the mean absolute error (MAE) and root mean square error (RMSE). Despite the significant progress achieved by satellite MVS, they still face bottlenecks in accurately computing depth. As shown in Fig. 1, the existence of these bottlenecks primarily

Received 19 April 2024; revised 19 July 2024 and 20 August 2024; accepted 7 September 2024. Date of publication 24 September 2024; date of current version 7 October 2024. This work was supported in part by the Science and Technology Planning Project of Guizhou Province, Department of Science and Technology of Guizhou Province, China under Grant [2023]159, in part by the Guizhou Province Higher Education Engineering Research Center under Grant [2023]41, and in part by the Natural Science Research Project of Guizhou Provincial Department of Education, China under Grant QianJiaoJi[2022] 029 and Grant QianJiaoHeKY[2021]022. *(Corresponding authors: Wenting Li; Yongjun Zhang.)*

Ziyang Chen and Yongjun Zhang are with the College of Computer Science, the State Key Laboratory of Public Big Data, Institute of Artificial Intelligence, Remote Sensing Satellite Image Big Data Innovation Center of Guizhou Province, Guizhou University, Guiyang 550025, China (e-mail: ziyangchen2000@gmail.com; zyj6667@126.com).

Wenting Li is with the School of Information Engineering, Guizhou University of Commerce, Guizhou 550023, China (e-mail: 201520274@gzcc.edu.cn).

Zhongwei Cui is with the School of Mathematics and Big Data, Guizhou Education University, Guizhou 550024, China (e-mail: zhongweicui@gznc.edu.cn).

Code will be available at https://github.com/ZYangChen/DC-SatMVS.

Digital Object Identifier 10.1109/JSTARS.2024.3457616

arises from two aspects: 1) misjudgment of depth resulting from a focus solely on local representations. Convolutional-based MVS architectures struggle to capture long-range information, which is crucial for satellite MVS. In contrast to general MVS scenes [12], satellite images exhibit locally similar textures, and the limited receptive field makes it challenging for existing algorithms to accurately estimate depth by incorporating the similarity of long-range information; 2) semantic confusion due to overconsideration of color characteristics. The loss functions commonly used in general MVS have limited efficacy in satellite MVS scenarios. Remote sensing images often share similar color information, rendering the supervision of satellite MVS training based solely on RGB values inefficient. Training satellite MVS with information unrelated to color, such as feature distribution [13], [14], is a more reasonable approach.

In response to the above challenges, we propose the Distribution Contrast Network for Satellite MVS (DC-SatMVS), a novel satellite stereo matching network. Specifically, we designed separate CNN and ViT branches for the learning of near-field and far-field features, respectively. In addition, we introduce the Distribution Contrast Loss (DCL), a contrastive loss calculated based on feature distributions. This loss function facilitates more rational supervision by considering network representations unrelated to color.

Overall, our contributions can be summarized as follows.

1) A new paradigm for satellite MVS that incorporates global information and feature distribution considerations is introduced.
2) We propose a dual-branch feature extractor that allows the network to capture local and global information simultaneously.
3) In order to strengthen the supervisory role of noncolor features on the network, we designed the DCL.
4) Our method achieves state-of-the-art (SOTA) results in WHU-TLC dataset, exhibits a 18.14% reduction in RMSE compared to Sat-MVSF.
5) Our design is also applicable to general MVS scenarios. In comparison to CasMVSNet, we achieves a notable 26.35% reduction in completeness errors.

II. Related Work

A. Manual Methods for Surface Depth Estimation

In the past few years, surface depth estimation of the Earth has mainly been achieved through manual geometric methods. Traditional manual methods can be broadly categorized into two main types.

The first type is based on the epipolar geometry of satellite images. An example of this approach is the RPC Stereo Processor (RSP) [15]. In this type, stereo images are first rectified according to the RPC [16] model, and then manual stereo matching algorithms such as Semiglobal Matching [17] are used to estimate disparities. Finally, the disparity map is converted into 3-D points in the world coordinate system.

The second type involves fitting a complex RPC model into a pinhole model for a small area and then, using stereo-matched pipelines for reconstruction. An example of this type is the Satellite Stereo Pipeline [18], which adjusts such stereo matching algorithms into the COLMAP framework for surface depth estimation of the Earth [6].

B. Learning-Based MVS

With the development of deep learning, learning-based MVS methods [8], [19] have demonstrated outstanding performance. As representative works of learning-based MVS, MVSNet [8], MVSNet++ [20], and P-MVSNet [21] employ a series of 3-D convolutions to regularize the cost volume. This approach requires a significant amount of GPU memory.

To address this limitation, a mainstream approach is to use recursive regularization methods to update depth estimations iteratively. For example, R-MVSNet [22] processed cost volumes at different depths using recursive regularization. All these methods were originally developed for natural images. RED-Net [10] extended the regularization method based on Convolutional Gated Recurrent Units. Recently, SOTA methods have introduced multistage cost volume construction approaches, such as CasMVSNet [7] and UCS-Net [9]. These methods have achieved outstanding performance, even when extended to depth estimation from satellite images, demonstrating strong generalization capabilities.

C. Learning-Based MVS for Surface Depth Estimation

To enhance the performance of MVS methods in satellite image depth estimation, Sat-MVS [3] introduces a rigorous RPC warping module as a plugin to existing MVS methods. Experiments demonstrate that this plug-in exhibits outstanding performance within various MVS frameworks [7], [9], [10].

Building upon this, a universal deep learning-based framework, named Sat-MVSF [4], is proposed for depth estimation from multiview optical satellite images of the Earth's surface. The framework consists of a complete processing pipeline, including preprocessing, a MVS network specialized for satellite images (Sat-MVSNet), and postprocessing. Preprocessing involves the geometric and radiometric configuration of the multiview images, as well as their cropping. The cropped multiview patches are then input into Sat-MVSNet, which performs depth feature extraction, RPC distortion, pyramid cost volume construction, regularization, and regression to obtain the height map. Error-prone matches are subsequently filtered out, and a Digital Surface Model is generated in the postprocessing stage. This approach achieves SOTA performance on the remote sensing image dataset WHU-TLC [3].

However, the aforementioned method still employs the common loss functions used in MVS, which exhibit limitations in performance on remote sensing datasets. This is because remote sensing images often have low color contrast, and the loss functions designed for remote sensing images should attenuate the influence of color factors on matching judgments. In addition, existing MVS methods still utilize CNN architectures, which restrict the ability of MVS to learn global information.

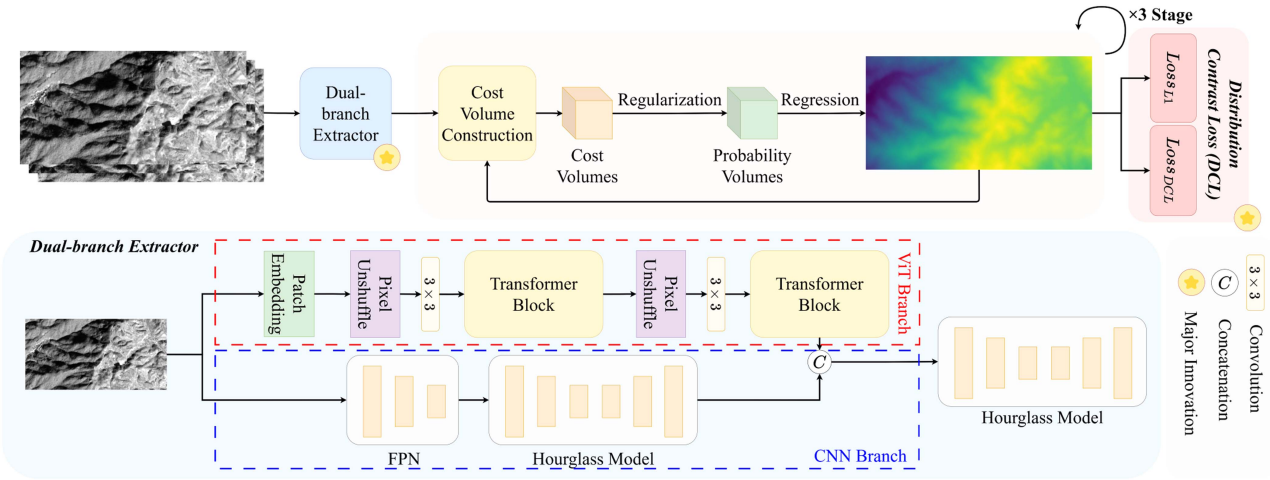

Fig. 2. Illustration of our method. We propose the ViT branch and CNN branch for feature extraction. The details of this feature extractor are illustrated in the second row. The obtained features follow the workflow of Sat-MVS (RED-Net) [3] to produce an estimated depth. Additionally, addressing the characteristic of minor color variations in remote sensing images, we introduce the Distribution Contrast Loss (DCL). The star means our major contributions.

III. Method

In this section, we first describe the overall architecture of our DC-SatMVS (Section III-A). Then, in Section III-B.B we present details of Dual-branch Extractor in DC-SatMVS. Finally, in Section III-D, we explain the DCL proposed by us.

A. General Architecture Overview

We proposed a novel satellite MVS Network (DC-SatMVS). Inspired by Sat-MVS [3], we construct a cost volume based on RPC warping and incorporate SOTA methods [3], [7], [10] from the MVS domain for cost map regularization and regression module design. Departing from the commonly used Feature Pyramid Network (FPN) [23] and U-Net [24] architectures in existing satellite MVS methods [3], [4], [10], we propose a paradigm that fuses ViT and CNN to enhance long-range information while preserving local semantics. In addition, considering the minimal color variation in remote sensing images, we introduce a Loss function, i.e., DCL, that focuses on frequency domain feature distribution. The overall workflow is illustrated in Fig. 2.

B. Dual-Branch Extractor

Existing satellite MVS methods [3], [4], [10] commonly adopt the CNN paradigm [11] for feature extraction. However, the limited receptive field of the convolutional structure hinders the acquisition of long-range features. Moreover, the minimal local content variation in remote sensing images poses a bottleneck for CNNs in extracting features from such imagery. We contend that, for remote sensing images, long-range information is equally crucial. This is because long-range interactions allow the network to recognise terrain relationships in the images, helping the network to achieve more accurate depth estimation by discovering general terrain patterns. Overall, CNN-based satellite MVS can only rely on local information for estimation, resulting in limited knowledge acquisition.

To complement the limited knowledge of CNNs, we propose a novel dual-branch Extractor. It captures both long-range semantics and short-range semantics through ViT and CNN branches, respectively. As illustrated in Fig. 2, our feature extractor presents a paradigm distinct from previous parallel CNN and ViT approaches. Specifically, the ViT branch encodes the input image into an eight-channel feature through Patch Embedding, mapping the feature to a higher-dimensional space via pixel unshuffle and convolution. To further capture long-range relationships, we construct the Transformer Block as depicted in Fig. 3. Let $n = H \times W$, due to the standard ViT design leading to $O(n^2)$ computational complexity, which is expensive for MVS tasks, we draw inspiration from Restormer [25] and design a matrix multiplication along the channel dimension. We are the first to apply this design to satellite MVS, leveraging it for long-range information interaction. However, this does not imply the abandonment of local information; we also design the CNN branch. The CNN branch acquires multiscale features related to local details through FPN and the Hourglass Model, inspired by [3] and [10]. Finally, we concatenate long-range features from the ViT branch and local features from the CNN branch to get the feature map.

It must be acknowledged that combining ViT [26] and CNN [11] is a strategy widely employed in various applications. Incorporating the CNN component into the Multihead Self-attention module and the feed forward propagation module is a common practice [27], [28], [29]. Our ViT branch similarly adopts this design strategy. However, the similarity of ideas does not mean that dual-branch extractors are exactly the same as the methods used in this type of strategy. These strategies still have the following problems: 1) combined with ViT is expensive. Attention has a square-level complexity. Let $n = H \times W$, the computational complexity of the attention matrix will reach $O(n^2)$ level. This means that the combination of ViT will significantly increase the cost of computing; 2) Weak local learning ability of feedforward networks. The vast majority of studies retain the use of MLP for forward propagation [27], [28], [29].

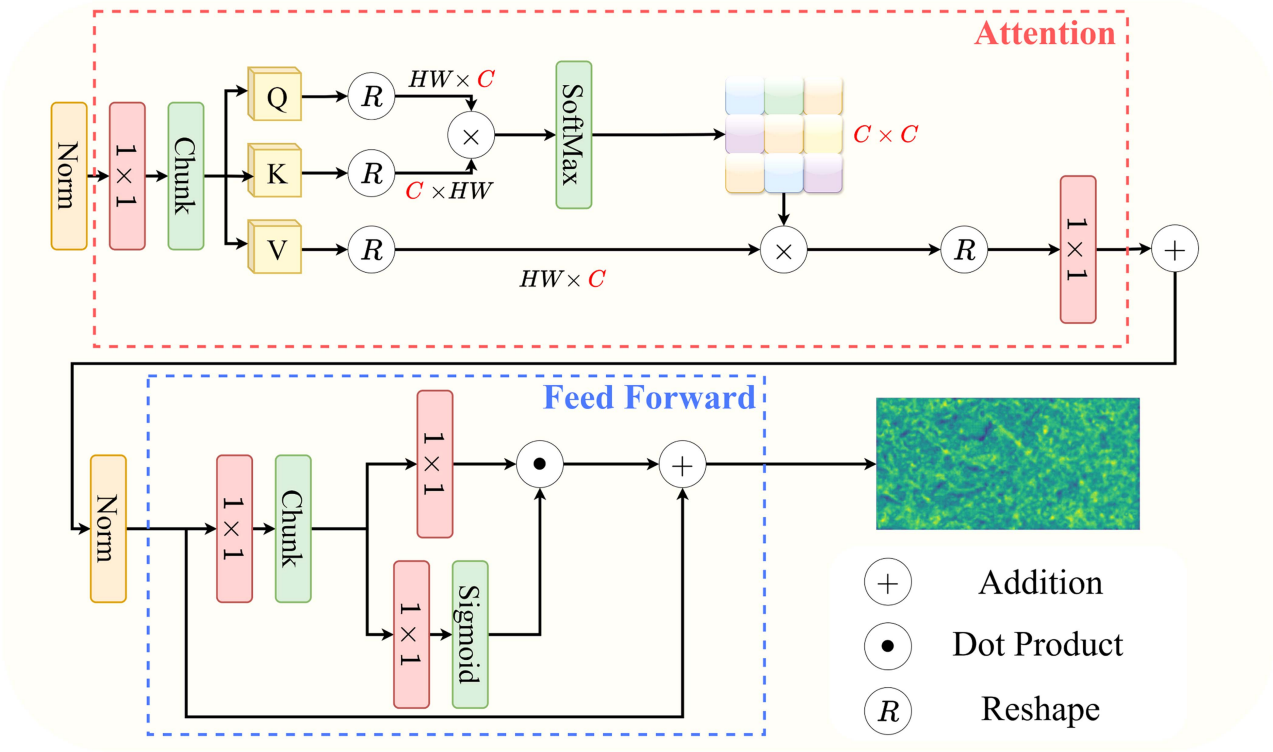

Fig. 3. Details of Transformer Block in our Dual-branch Extractor. The symbol R means $reshape$. 1×1 denotes CNN layers with kernel size set to 1.

The patch-wise learning paradigm is unfavorable for intrablock feature learning. It is difficult to adapt to the complex features of remotely sensed imagery.

To solve the above limitation, our transformer module has the following improvements: 1) transpose the Q and K matrices, and then perform matrix multiplication. In this way, we control the time complexity of the attention matrix computation to $O(C^2)$. Since the number of feature channels $C << n$, this approach actually achieves $O(n)$ linear complexity. A powerful attention head is achieved by interacting only the attention of the channels; 2) pixel-wise feature propagation with convolution and gating. We eliminate the linear layer + activation function paradigm and use convolution and gating for pixel-by-pixel feature updating. In addition, objects in remote sensing images vary significantly in scale. A CNN branch is still designed to capture multiscale local semantics and help the ViT branch achieve better performance.

C. Cost Volume Construction

Cost metrics of DC-SatMVS follows [8], [30]. We calculate the variance of the element values at corresponding positions in the feature volumes to form a single cost volume. A three-stage cost volume is built in a cascaded manner [3], [7] based on this single volume, allowing features to be matched from coarse to fine. Regularization and regression of probability volumes also follows [3].

D. Distribution Contrast Loss (DCL)

Existing satellite MVS methods [3], [4] commonly align with conventional MVS approaches [7], [8], employing Loss_{L1} or $\text{Loss}_{\text{Smooth}^{\sim}L1}$ for loss computation. However, this type of loss calculation may not be the most rational choice. The standard MVS computation inherently involves the contrast of color information, which is not applicable to remote sensing images with minimal color variation. To emphasize features unrelated to color, we propose a DCL that focuses on the frequency domain distribution.

Following the workflow outlined in Fig. 2, we can obtain the estimated depth d at each stage. Next, the result obtained is shifted by a two-dimensional discrete fourier transform f and the operation $fftshift$. $fftshift$ moves the zero-frequency component to the center of the array. This increases the symmetry of the spectrum. In this way, we learn about the distribution of depth information in the frequency domain. Furthermore, we employ a normalization strategy to stabilize the training process. Both the estimated depth and ground truth undergo this transformation into the frequency domain, a process expressed by

$$\begin{aligned} d^f &= fftshift(f(d)) \\ d^f_{\text{norm}} &= \frac{d^f - \min(d^f)}{\max(d^f) - \min(d^f)} \end{aligned} \tag{1}$$

where min and max refer to taking the minimum and maximum values, respectively. In frequency space, a more intuitive comparison of high and low frequency information allows us to divide this information into k groups, ensuring that the frequency distribution within each group approximates the frequency distribution of the ground truth. We quantify this approximation using (4), employing the Kullback–Leibler (KL) divergence. The KL divergence measures the closeness of two distributions:

the smaller the KL divergence, the closer the distributions; conversely, a larger KL divergence indicates greater dissimilarity. This characteristic allows us to incorporate it as part of the loss function. Finally, we utilize a piecewise function, as depicted in (5), to compress the scale of this relationship, promoting smoother training and obtaining the value of the DCL.

$$d^{f\prime} = \log\left(\text{SoftMax}\left(d^{f}_{\text{norm}}\right)\right) \tag{2}$$

$$d^{f\prime}_{gt} = \text{SoftMax}\left(d^{f}_{gt\_\text{norm}}\right) \tag{3}$$

$$kl(d^{f}, d^{f}_{gt}) = \sum_{j=1}^{k}\left(KL\left(d^{f\prime}, d^{f\prime}_{gt}\right)\right) \tag{4}$$

$$\text{Loss}_{\text{DCL}}(d^{f}, d^{f}_{gt}) = \begin{cases} \log\left(kl\left(d^{f}, d^{f}_{gt}\right)\right), & kl > 1, \\ 0, & kl \leq 1. \end{cases} \tag{5}$$

The final loss function for the DCL is obtained according to Formula 6, where we integrate considerations for both the differences in frequency domain distribution and specific numerical disparities.

$$\begin{aligned} \text{Loss} = {} & \gamma_1 \times \text{Loss}_{\text{Smooth}L1}\left(d, d_{\text{gt}}\right) \\ & + \gamma_2 \times \text{Loss}_{\text{DCL}}\left(d^{f}, d^{f}_{\text{gt}}\right) \end{aligned} \tag{6}$$

where γ_1 and γ_2 are factors of loss. We set $\gamma_1 = 0.8$ and $\gamma_2 = 0.2$, respectively.

In other fields of study, both Zheng et al. [31] and Zhang et al. [32] have proposed methods for calculating loss based on distribution. Zheng et al.'s [31] approach involves using a larger number of samples as anchors and obtaining feature contrasts through a VGG network [33]. However, this loss has relatively poor interpretability and is computationally expensive. On the other hand, the approach of Zhang et al. [32] can lead to negative results when the distributions are close together, which can affect the robustness of the training. Our DCL provides a novel perspective for this type of distribution-based loss calculation.

IV. Experiments

This section presents the efficacy of the DC-SatMVS. Section IV-A outlines the experimental setup, while Section IV-B showcases the outcomes of the experiments. Section IV-C illustrates the contribution of each module through ablation experiments. In addition, the approach can be applied to generic MVS scenarios, with the corresponding evidence presented in Section IV-D.

A. Experimental Setup

Due to the challenging nature of collecting MVS datasets, the availability of existing remote sensing MVS datasets is relatively limited. We chose to validate our approach using the WHU-TLC [3] dataset due to its novelty. In addition, this dataset has a higher level of parameter openness, and numerous approaches have been validated on it. This facilitates comparisons with a broader range of algorithms. The framework was implemented in PyTorch and trained on single NVIDIA A6000. The hyperparameters follow [3], in the training phase the batch size was set to 1 and RMSprop was chosen as the optimizer. Our network is trained for 15 epochs with an initial learning rate of 0.001, and are downscaled by a factor of 2 after the 10th epoch.

In order to thoroughly assess the efficacy of our proposed method, we expand our evaluation to encompass traditional MVS scene datasets. Specifically, we choose the widely recognized DTU dataset [12] for training and testing purposes. Following the common practice, we train our network on the DTU training set and evaluate it on the DTU evaluation set while adopting the same data split and view selection as defined in [7] for a fair comparison. The number of input images is set to N = 5 with a resolution of 640 × 512 for the DTU. We trained on the DTU with the Adam optimizer for 16 epochs from a start learning rate of 0.001 on 4 NVIDIA Tesla T4 GPUs.

B. Experimental Results

To assess the effectiveness of our method in satellite MVS scenarios, we compare our results against both manual methods and learning-based approaches. Quantitative results are presented in Table I. The MAE, RMSE, and L1 distance error metrics were employed. They are commonly utilized [3], [4], [10] for the estimation of depth in remote sensing images. MAE represents the mean absolute error between the predicted value and the ground truth (GT), RMSE represents the sample standard deviation of the difference between the predicted value and the GT, and L1 distance error represents the error between the predicted value and the GT under the permissible deviation threshold.

Compared to the best-performing Sat-MVSF [4] in terms of MAE and RMSE metrics, we achieve a reduction of **5.17%** and **18.14%**, respectively. Moreover, in comparison to SatMVS (RED-Net) [3], [10], which exhibits the best accuracy within 2.5 meters, our method elevates the SOTA level of accuracy by **6.22%**. Visual comparisons are depicted in Fig. 4. The red-framed areas in the figure illustrate that existing methods tend to incorrectly estimate the depth of an entire area as a uniform value. Moreover, these methods tend to estimate the depth distance as being closer than the true value. From the visualizations, it is evident that SOTA methods still exhibit noticeable estimation errors, while our method consistently ensures a reasonable estimation outcome.

We also perform qualitative analyses in different types of regions. The visualization results, as shown in Fig. 5, illustrate the ability of DC-SatMVS to achieve clear results. DC-SatMVS works effectively under conditions of degradation and various land cover types.

C. Ablation Study

To validate the effectiveness of each design, we conducted ablation experiments on the WHU-TLC dataset [3], and the role of each module is outlined in Table II. No. 1 refers to the validation results of our baseline method. In the No.2 experiment, we replaced the $\text{Loss}_{\text{Smooth}L1}$ used in SatMVS (RED-Net) [3] with our proposed DCL Loss, resulting in a 7.38% improvement in accuracy within 2.5 m. In the No. 4 experiment, we replaced

TABLE I
QUANTITATIVE RESULTS OF THE DIFFERENT METHODS ON THE WHU-TLC DATASET

Methods	Year	MAE (m)↓	RMSE (m)↓	<2.5m (%)↑	<7.5m (%)↑	Runtime↓
Adapted COLMAP [6]	2019	2.227	5.291	73.35	96.00	77 min 27 s
CATALYST	2021	3.454	7.939	52.31	82.52	3 min 48 s
ArcGIS	2022	4.607	10.689	48.88	77.71	6 min 49 s
CasMVSNet [7]	2020	2.031	4.351	77.39	96.53	4 min 02 s
RED-Net [10]	2020	2.171	4.514	74.13	95.91	9 min15 s
UCS-Net [9]	2020	2.039	4.084	76.40	96.66	**3 min47 s**
SatMVS(RED-Net) [3]	2021	1.945	4.070	77.93	96.59	13 min52 s
SatMVS(CasMVSNet) [3]	2021	2.020	3.841	76.79	96.73	12 min20 s
SatMVS(UCS-Net) [3]	2021	2.026	3.921	77.01	96.54	13 min17 s
Sat-MVSF [4]	2023	1.895	3.654	64.82	80.05	5 min52 s
DC-SatMVS	Ours	**1.797** (−5.17%)	**2.991** (−18.14%)	**84.15** (+6.22%)	**97.07** (+0.34%)	19 min36 s

Above the central horizontal line are manual methods, below are learning-based methods. The best performing result is bold, the second best performing result is underlined, and the percentage sign indicates the improvement of the best performance compared to the second best performance. The runtime of our method is obtained through inference on single NIVIDA Tesla A100 GPU.

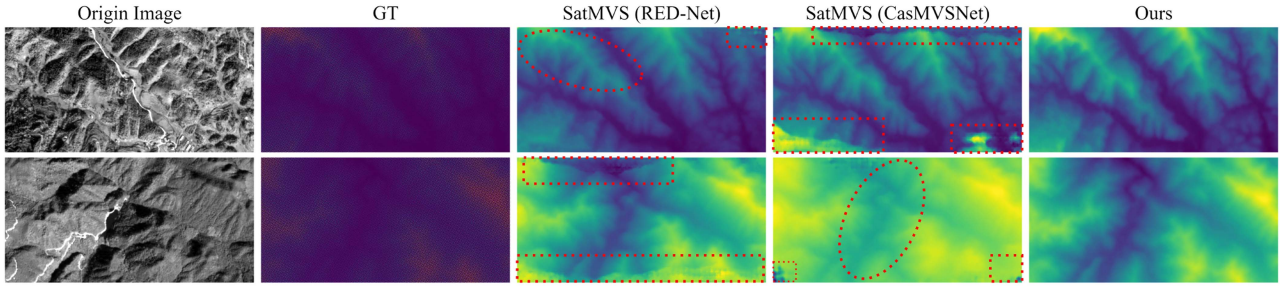

Fig. 4. Visual comparisons on the WHU-TLC dataset. Existing methods have a tendency to incorrectly estimate the depth of an entire area as a similar value when determining depth. In addition, they also tend to estimate the distance of the depth closer than the true value. We solve the above problems by adding knowledge of long distances and improving the training of noncolor features.

TABLE II
ABLATION RESULTS ON THE WHU-TLC TEST DATASET

No.	Feature extractor	Loss	MAE (m)↓	<2.5m (%)↑	<7.5m (%)↑	Runtime↓
1	FPN [3]	[3]	1.945	77.93	96.59	18 min 18 s
2	FPN [3]	DCL	1.928	83.68	96.86	**18 min 04 s**
3	SWT [34]	[3]	1.933	80.71	96.66	31 min 45 s
4	DBE	[3]	1.808	83.89	97.02	19 min 12 s
5	SWT [34]	DCL	1.882	84.02	96.96	31 min 32 s
6	DBE	DCL	**1.797**	**84.15**	**97.07**	19 min 36 s

Our baseline is SatMVS (RED-Net) [3]. To emphasize the modules innovated in our approach, module names highlighted in red text represent our contributions. DBE denotes our Dual-branch extractor. FPN denotes the feature pyramid network [3] structure. SWT denotes the Swin transformer [34] structure. All runtime measurements presented below are obtained through inference on single Nivida Tesla A100 GPU.

TABLE III
PARAMETER SENSITIVE ANALYSIS IN THE WEIGHT OF THE DISTRIBUTION CONTRAST LOSS

No.	γ_1	γ_2	MAE (m)↓	<2.5 m (%)↑	<7.5 m (%)↑
1	0.9	0.1	1.812	83.87	96.62
2	0.8	0.2	**1.797**	84.15	**97.07**
3	0.7	0.3	1.804	**84.61**	97.05
4	0.6	0.4	1.809	83.91	96.95

The best performing result is bold.

the FPN used in SatMVS (RED-Net) with our proposed Dual-branch Extractor, achieving a 7.04% reduction in MAE. Experiment No. 6 simultaneously improved the feature extraction and loss calculation methods of the baseline method [3], producing SOTA results.

Sensitivity analysis of hyperparameters in the weighted loss function is presented in Table III. $\gamma_1 = 0.8$ and $\gamma_2 = 0.2$ are set for DCL, this set of parameters gives the best results. Possibly due to device differences, we did not achieve similar inference times to SatMVS [3] in its origin paper. This indicates that there is still room for optimization in terms of inference time for our method. In addition, it is acknowledged that the feature extractor may slow down inference time, but given the relatively low real-time inference demand for depth estimation in satellite imagery, the acceptable tradeoff for improved accuracy is justified. DCL appears only as a loss function and theoretically does not impact inference time. The observed runtime differences are likely due to variations in device usage, resulting in some degree of error. We also try to replace the feature extractor with the Swin Transformer [34] architecture, because it is a commonly used transformer architecture that serves as a strong baseline for improved ViT. The experiments in No. 3 and No. 5 show that Swin Transformer can achieve better performance than FPN. However, this classical transformer architecture is not as effective as our DBE. In addition, using Swin Transformer to extract features significantly increases the computation time due to the existence of attention-squared level computational

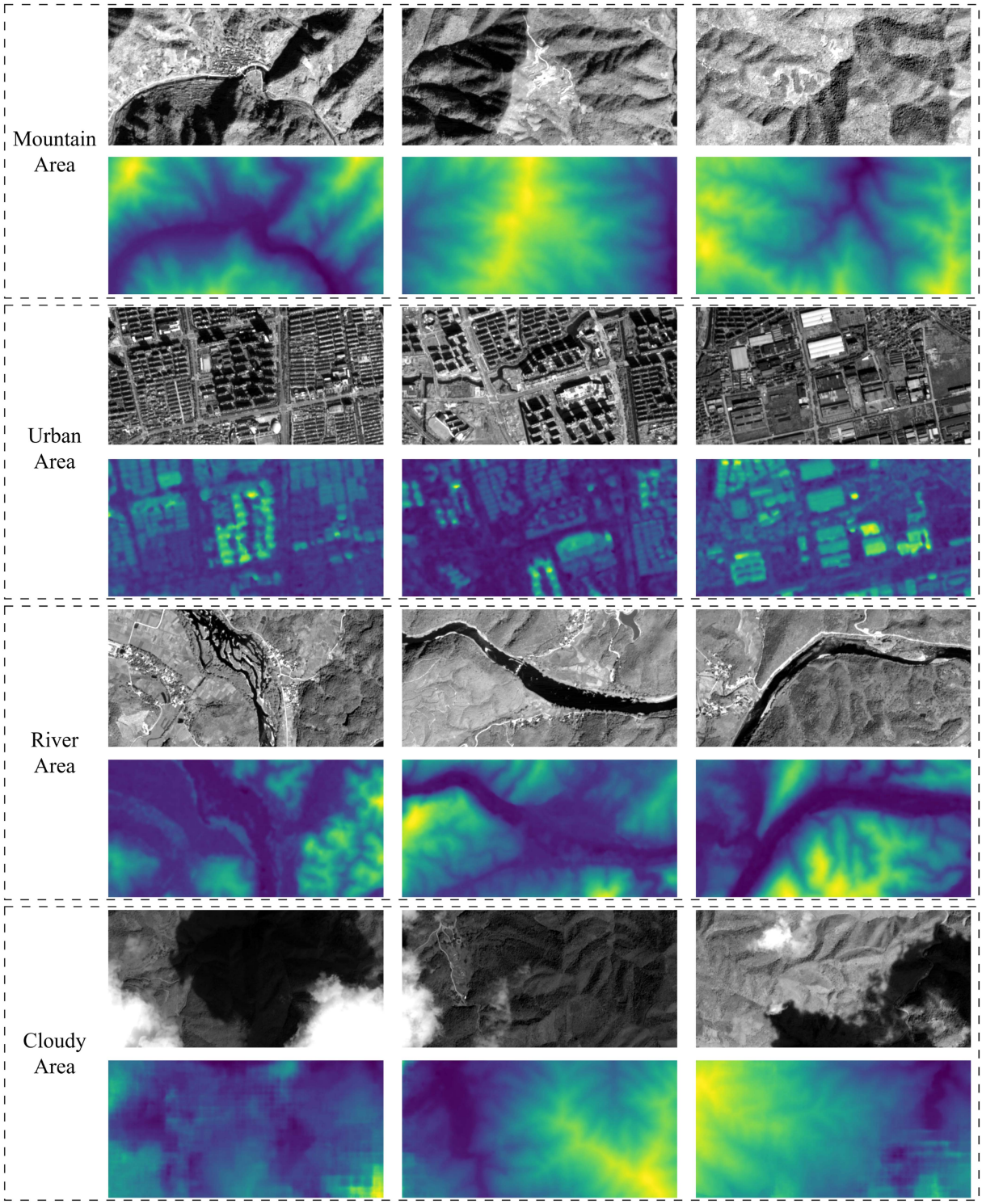

Fig. 5. Visualization of the results achieved by DC-SatMVS under different scenarios.

complexity. In summary, we can conclude that each of our designs proves effective for satellite Multi-view Stereo.

D. Cross-Dataset Generalization

To validate the effectiveness of our design, we transferred our approach to commonly used MVS scenes and achieved SOTA results. The numerical results and visualizations of the method are presented in Table IV and Fig. 6. Compared to CasMVS-Net [7], we, respectively, achieved reductions of 10.42% and 26.35% in the Overall performance and Completeness metrics, which are both error-related indicators. In comparison to the advanced IGEV-MVS [38], our method lowered these metrics by 1.85% and 5.38%. Even when compared to the latest

TABLE IV
PERFORMANCE ON THE DTU DATASET

Methods	Publish	**Ove.**(mm)↓	Acc.(mm)↓	Comp.(mm)↓
MVSNet [8]	ECCV2018	0.462	0.396	0.527
CasMVSNet [7]	CVPR2020	0.355	**0.296**	0.406
UCS-Net [9]	CVPR2020	0.344	0.338	0.349
IterMVS [35]	CVPR2022	0.363	0.373	0.354
Vis-MVSNet [36]	IJCV2023	0.365	0.369	0.361
DispMVS [37]	AAAI2023	0.339	0.354	0.324
IGEV-MVS [38]	CVPR2023	0.324	0.331	<u>0.316</u>
MoCha-MVS [30]	CVPR2024	<u>0.319</u>	<u>0.314</u>	0.325
DC-MVS	Ours	**0.318**	0.337	**0.299**

Ove. means overall performance, Acc. means accuracy, and comp. Means completeness. The smaller the better for each indicator. The best performing result is bold, the second best performing result is underlined.

Fig. 6. Visualization of partial scenes on the DTU dataset. We reconstructed these scenes based on the estimated depth of the point clouds.

algorithm MoCha-MVS [30], our DC-MVS still demonstrates superior performance. We have to acknowledge that our method does not exhibit superior performance in terms of *accuracy* metrics compared to SOTA methods. This phenomenon might be attributed to differences in features between remote sensing images and conventional images, resulting in a certain level of absolute position estimation offset in point clouds. However, this does not imply that our design is unsuitable for MVS scenes. According to the quantitative assessments in Table IV, our design significantly enhances the completeness of reconstruction. The incorporation of neighboring points compensates for absolute point cloud deviations, achieving a lower overall error compared to MVS reconstruction methods specialized for conventional scenes. This phenomenon highlights the importance of long-range capabilities and noncolor features in modeling completeness, even though they may compromise the absolute accuracy of the point cloud to some extent.

V. CONCLUSION

In this article, we propose the Dual-branch Extractor with DCL, a novel method for depth estimation in satellite MVS. The Dual-branch Extractor, introduced as a novel feature extractor for satellite MVS, overcomes the limitations of existing MVS methods in capturing long-range relationships. In addition, we introduce the DCL, a loss function focused on frequency domain distribution, reducing the emphasis on color in conventional MVS methods and enhancing training efficiency.

The objective of this process is to develop a more efficient depth estimation scheme for purely visual remote sensing. Experimental results demonstrate that our approach can be effectively applied to the task of depth estimation in multiview remote sensing images, achieving SOTA results on multiview remote sensing datasets. Furthermore, the framework exhibits exemplary generalization performance in generic MVS scenarios. In the future, we plan to extend these designs to other remote sensing image processing tasks.

REFERENCES

[1] L. Moya, E. Mas, and S. Koshimura, "Sparse representation-based inundation depth estimation using SAR data and digital elevation model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9062–9072, 2022.

[2] D. Li, G. Shi, W. Kong, S. Wang, and Y. Chen, "A leaf segmentation and phenotypic feature extraction framework for multiview stereo plant point clouds," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2321–2336, 2020.

[3] J. Gao, J. Liu, and S. Ji, "Rational polynomial camera model warping for deep learning based satellite multi-view stereo matching," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 6148–6157.

[4] J. Gao, J. Liu, and S. Ji, "A general deep learning based framework for 3D reconstruction from multi-view stereo satellite images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 195, pp. 446–461, 2023.

[5] Y. Lu, Y. Zhang, Z. Cui, W. Long, and Z. Chen, "Multi-dimensional manifolds consistency regularization for semi-supervised remote sensing semantic segmentation," *Knowl.-Based Syst.*, vol. 299, 2024, Art. no. 112032.

[6] K. Zhang, N. Snavely, and J. Sun, "Leveraging vision reconstruction pipelines for satellite imagery," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 2139–2148.

[7] 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 Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2495–2504.

[8] 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.

[9] S. Cheng et al., "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2524–2534.

[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 Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6050–6059.
[11] Y. LeCun et al., "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.
[12] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *Int. J. Comput. Vis.*, vol. 120, pp. 153–168, 2016.
[13] Y. Zhang, M. Li, R. Li, K. Jia, and L. Zhang, "Exact feature distribution matching for arbitrary style transfer and domain generalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8035–8045.
[14] Y. Wu, W. Li, Z. Chen, H. Wen, Z. Cui, and Y. Zhang, "Distribution-decouple learning network: An innovative approach for single image dehazing with spatial and frequency decoupling," *The Vis. Comput.*, Springer, vol. early access, pp. 1–16, 2024.
[15] R. Qin, "RPC stereo processor (RSP)–A software package for digital surface model and orthophoto generation from satellite stereo imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 3, pp. 77–82, 2016.
[16] M. Wang, F. Hu, and J. Li, "Epipolar resampling of linear pushbroom satellite imagery by a new epipolarity model," *ISPRS J. Photogrammetry Remote Sens.*, vol. 66, no. 3, pp. 347–355, 2011.
[17] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, IEEE, 2005, pp. 807–814.
[18] C. De Franchis, E. Meinhardt-Llopis, J. Michel, J.-M. Morel, and G. Facciolo, "An automatic and modular stereo pipeline for pushbroom images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 2, pp. 49–56, 2014.
[19] Z. Chen et al., "Feature distribution normalization network for multi-view stereo," *The Vis. Comput*, Springer, vol. early access, pp. 1–13, 2024.
[20] P.-H. Chen, H.-C. Yang, K.-W. Chen, and Y.-S. Chen, "MVSNet: Learning depth-based attention pyramid features for multi-view stereo," *IEEE Trans. Image Process.*, vol. 29, pp. 7261–7273, 2020.
[21] 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. Int. Conf. Comput. Vis.*, 2019, pp. 10452–10461.
[22] 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 Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5525–5534.
[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.
[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Computer-assisted Intervention*, Springer, 2015, pp. 234–241.
[25] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5728–5739.
[26] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2020.
[27] H. Yan, Z. Li, W. Li, C. Wang, M. Wu, and C. Zhang, "ConTNet: Why not use convolution and transformer at the same time?" 2021, *arXiv:2104.13497*.
[28] Z. Peng et al., "Conformer: Local features coupling global representations for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 367–376.
[29] Y. Chen et al., "Mobile-former: Bridging MobileNet and transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5270–5279.
[30] Z. Chen et al., "MoCha-stereo: Motif channel attention network for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 27768–27777.
[31] Y. Zheng, J. Zhan, S. He, J. Dong, and Y. Du, "Curricular contrastive regularization for physics-aware single image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5785–5794.
[32] Z. Zhang, R. Peng, Y. Hu, and R. Wang, "GeoMVSNet: Learning multi-view stereo with geometry perception," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 21508–21518.
[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.
[34] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[35] S. Wang, B. Li, and Y. Dai, "Efficient multi-view stereo by iterative dynamic cost volume," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8655–8664.
[36] J. Zhang, S. Li, Z. Luo, T. Fang, and Y. Yao, "Vis-MVSNet: Visibility-aware multi-view stereo network," *Int. J. Comput. Vis.*, vol. 131, no. 1, pp. 199–214, 2023.
[37] Q. Yan, Q. Wang, K. Zhao, B. Li, X. Chu, and F. Deng, "Rethinking disparity: A depth range free multi-view stereo based on disparity," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 3091–3099.
[38] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 21919–21928.

Ziyang Chen (Student Member, IEEE) is currently working toward the master's degree in computer science and technology with the College of Computer Science and Technology, Guizhou University, Guiyang, China. He is supervised by Prof. Yongjun Zhang.

His research is also funded by Prof. Wenting Li. He has authored or coauthored several papers in journals and conferences such as Conference on Computer Vision and Pattern Recognition (CVPR). His research interests include stereo matching, remote sensing, and intelligent transportation.

He is a Reviewer for *Knowledge-based Systems* and *Electronic Letters*.

Wenting Li received the B.Sc. degree in computer science and technology from Zhengzhou University of Aeronautics, Zhengzhou, China, in 2006, the M.Sc. degree in computer science from GuiZhou University, Guiyang, China, in 2010, and the Ph.D. degree in computer technology and application from the Macau University of Science and Technology, Macau, China, in 2017.

She is a Doctor of computer technology and application and Professor with the Guizhou University of Commerce, Guiyang, China, since 2018. Her research interests include intelligent transportation, computer vision, and data analysis.

Zhongwei Cui received the master's degree in computer application technology from Guizhou University, Guiyang, China, in 2008.

He is currently Associate Professor with the School of Mathematics and Big Data, Guizhou Education University, Guiyang, China, since 2013. His research interests include remote sensing and wireless networks.

Yongjun Zhang (Member, IEEE) received the M.Sc. and Ph.D. degrees in software engineering from Guizhou University, Guiyang, China, in 2010 and 2015, respectively.

From 2012 to 2015, he is a joint training doctoral student of Peking University, Beijing, China, and Guizhou University, studying in the key laboratory of integrated microsystems of Peking University Shenzhen Graduate School, Shenzhen, China. He is currently an Associate Professor with Guizhou University. His research interests include the intelligent image algorithms of computer vision, such as scene target detection, remote sensing, stereo matching, and low-level vision.