A Multitask Network for Multiview Stereo Reconstruction: When Semantic Consistency-Based Clustering Meets Depth Estimation Optimization

Xin Huang, *Senior Member, IEEE*, Shulei Zhang, Jiayi Li, *Senior Member, IEEE*, and Leiguang Wang

***Abstract*— We propose a novel network for multiview stereo (MVS) reconstruction in the field of remote sensing, which considers clustering-based semantic consistency into depth estimation optimization, referred to as CSC-MVS. In this approach, high-level semantic information acquired from multiple views is used to construct semantic consistency and assist in guiding the optimization of the MVS network. Specifically, the nonnegative matrix factorization (NMF) branch and the deep spectral decomposition (DSD) branch are designed to generate local and global semantic guidance, respectively. We then propose an uncertainty multitask optimization method to adaptively combine matching and semantic metrics. The performance of CSC-MVS is evaluated on representative benchmarks, including the WHU TLC dataset and LuoJia-MVS dataset, demonstrating its effectiveness and generality across diverse remote sensing scenarios. Comprehensive experimental results show that our CSC-MVS significantly improves the performance of various MVS baseline networks and achieves notable accuracy in depth reconstruction. We also conduct ablation studies to validate the rationality of each component and sensitivity analysis to confirm the robustness and adaptability of our proposed method. The code is available at https://github.com/zsl-whu/csc-mvs.**

***Index Terms*— Clustering algorithm, deep learning, depth map-based multiview stereo (MVS) reconstruction, remote sensing imagery.**

I. Introduction

A. Background

MULTIVIEW stereo (MVS) aims to recover a dense representation of a 3-D scene by leveraging multiple overlapping images and calibrated cameras [1]. In recent years, with the advent of deep learning techniques, it has emerged as a prominent research topic in the intersection of computer vision and remote sensing [2], [3]. Particularly, deep learning methods have shown great potential in adapting to scenes with limited texture information or varying lighting conditions [4] while also alleviating the time-consuming and inefficient issues of traditional approaches [5], [6]. Thus, although reliance on training data is required, the introduction of deep learning-based MVS techniques for large-scale terrain reconstruction tasks in remote sensing holds important significance and research potential [7], [8].

B. Status

In the development of deep learning-based MVS techniques, benchmarks play an indispensable role [9], [10]. Differing from computer vision datasets that mainly consist of stable indoor scenes [11], [12], [13], remote sensing images are captured from a considerable distance, often measured in kilometers. The topographic variations within a single image frame in remote sensing datasets are much larger than those of desktop objects or specific landscapes [14]. This not only results in an obvious difference in spatial resolution between the two types of datasets but also leads to a much larger range of depth variations in remote sensing MVS datasets compared to computer vision datasets [6], [15], [16]. Consequently, the corresponding depth interval scale in remote sensing MVS datasets is several times greater than the latter (see Fig. 1). Therefore, directly applying MVS methods originating from the field of computer vision to large-scale remote sensing 3-D reconstruction poses notable challenges in terms of feasibility and accuracy [17].

In the 3-D scene reconstruction algorithms, the methods based on depth map estimation [18] have gained considerable attention due to their advantages over methods relying on voxel [19] or point cloud [20], which often suffer from memory consumption limitations [21], [22], [23]. MVSNet [4], the first end-to-end 3-D reconstruction deep learning algorithm based on depth map estimation, has further spurred research in this area. It comprises four key modules: 1) feature extraction, 2) 3-D cost volume construction, 3) cost volume regularization, and 4) depth estimation.Subsequent to the development of MVSNet, mainstream MVS networks can be categorized into two types: noncascade and cascade. Table I follows

Manuscript received 6 October 2023; revised 29 December 2023; accepted 23 February 2024. Date of publication 28 February 2024; date of current version 11 March 2024. This work was supported in part by the Major Scientific and Technological Projects of Yunnan Province under Grant 202202AD080010, in part by the National Natural Science Foundation of China under Grant 42071311 and Grant 42271328, and in part by the Special Fund of Hubei Luojia Laboratory under Grant 220100031. *(Corresponding author: Jiayi Li.)*

Xin Huang and Shulei Zhang are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: xhuang@whu.edu.cn; zhangshulei@whu.edu.cn).

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

Leiguang Wang is with the Institute of Big Data and Artificial Intelligence, Southwest Forestry University, Kunming 650024, China, and also with the Key Laboratory of State Forestry and Grassland Administration on Forestry and Ecological Big Data, Southwest Forestry University, Kunming 650024, China (e-mail: wlgbain@126.com).

Digital Object Identifier 10.1109/TGRS.2024.3371059

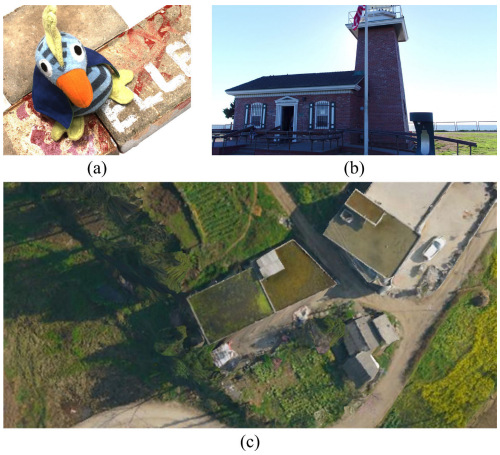


Fig. 1. Comparison of close-range dataset and remote sensing dataset. The sample reference images above are selected from (a) the DTU dataset [11], (b) the Tanks and Temples dataset [12], and (c) the LuoJia-MVS dataset [15]. When the number of depth planes is set to 192, the depth intervals for the close-range images in (a) and (b) are 2.65 and 4.86 mm, respectively, whereas the depth interval for the remote sensing data in (c) is 12.7 cm.

this grouping structure and summarizes the characteristics, advantages, and limitations of several state-of-the-art (SOTA) methods.

C. Motivations

Recent studies have basically followed the framework and primarily focused on enhancing the initial three modules [29]. Regarding depth estimation, R-MVSNet [5] introduced a shift from regression to multiclass classification, while UniMVS [28] proposed an approach to unify the advantages of regression and classification. Research emphasizing this crucial module, however, remains relatively scarce; moreover, although MVSNet-based algorithms have made progress in close-range scene reconstruction, their applicability in remote sensing, which involves diverse factors, such as varying resolutions, depth ranges, and scales, remains unexplored [15]. Additionally, despite existing improvements leading to an increase in model complexity and network size [8], [29], the optimization of depth estimation continues to rely exclusively on dense matching, presenting inherent challenges to algorithmic robustness. To tackle these challenges, this study proposes a general approach called "CSC-MVS," designed to robustly enhance the performance of MVS networks in remote sensing benchmarks without requiring additional data assistance.

Inspired by the research on photometric consistency filtering in depth estimation [5], it is believed that the abundant semantic information present in remote sensing images can provide valuable assistance in depth estimation [30], [31]. Photometric consistency assumes that corresponding points in different views should exhibit consistency in terms of brightness and texture [32]; however, this assumption is vulnerable to factors such as lighting variations, noise, and occlusions [33]. In contrast, semantics, which can represent a more intrinsic characteristic of images [34], have the potential to guide depth estimation in a more robust manner [35].

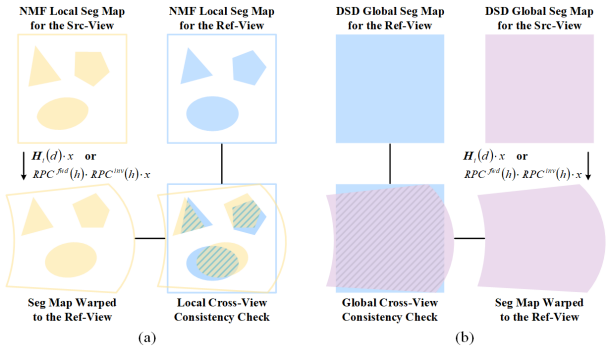


Fig. 2. Process for establishing semantic consistency and enhancing depth estimation using NMF and DSD. (a) Mechanism of NMF. (b) Mechanism of DSD. Blue denotes the semantic map of the reference view, while yellow and purple, respectively, indicate the local and global semantic maps from the source view. Discrete polygons and solid-colored blocks represent common local and global semantics. The shaded regions in the cross-view check images highlight areas where semantic correspondence is correct.

It is worth noting that under the guidance of semantic consistency, the construction of appropriate semantic features becomes necessary [36]. Nonetheless, directly relying on semantic labels for enhancement is computationally expensive due to the limitations of the remote sensing data scale [37], and incorporating an additional semantic segmentation network branch would further escalate the training cost. Additionally, how to integrate the constructed semantic features into the existing MVS algorithm framework is also an important aspect that requires consideration in this study. Given this context, our research proposes to use unsupervised clustering to deal with the above problems.

Fig. 2(a) and (b) illustrate the processes of nonnegative matrix factorization (NMF) and deep spectral decomposition (DSD) to achieve semantic consistency from local and global perspectives, respectively. Each approach begins by extracting local (or global) semantic maps from the feature matrix derived from multiview images. The semantic map generated from the source view is then warped to the reference view based on the predicted depth values. The similarity between the warped and the original reference semantic maps, denoted by the overlap of the two polygons in the shaded area, constitutes the desired semantic consistency. By enhancing the cross-view consistency, the predicted depth maps can be improved.

D. Contributions

The main contributions of this article are as follows.

1) In the context of 3-D reconstruction tasks in remote sensing, CSC-MVS, a semantic enhancement method, is designed for MVS networks. By combining unsupervised clustering algorithms with MVS networks, both local and global semantic features have been leveraged to provide reliable guidance for depth optimization.
2) To improve the depth estimation module of MVS networks, a multitask depth optimization objective has been proposed. In addition to dense matching, cross-view semantic consistency has been incorporated in the form of uncertainty multitask loss, providing benefits for depth estimation.

TABLE I
SUMMARY OF THE SEVEN SOTA SUPERVISED LEARNING (TRAINING DATA REQUIRED) MVS NETWORKS

Model Type	Model Name	Characteristics	Advantages	Limitations	Depth Estimation Mode
Non-Cascade	MVSNet [4]	Based on plane sweeping, depth is estimated in a manner like epipolar line search.	End-to-end 3D reconstruction is achieved through differentiable homography.	High memory consumption, unsuitable for large-scale scenes.	Regression
	R-MVSNet [5]	On the basis of MVSNet, a slice-wise recurrent network is sequentially employed for regularization.	Memory consumption is significantly reduced.	Inferior reconstruction accuracy and expensive processing time.	Classification
	D2HC-RMVSNet [24]	On the basis of R-MVSNet, a dense receptive expansion network is employed to extract multi-scale features, and the regularization approach is further refined into a hybrid recurrent network.	Reconstruction accuracy is improved through multi-scale feature aggregation and rich contextual information.	Complex network modules with an excessive size of parameters.	Classification
	AA-RMVSNet [25]	On the basis of R-MVSNet, deformable convolutional kernels are incorporated for multi-scale texture features, and an adaptive aggregation layer is introduced for cost volume construction.	Reconstruction quality of challenging areas like low texture and occlusions is enhanced.	Adaptive aggregation layers may lead to insufficient matching information and model underfitting.	Classification
Cascade	Cas-MVSNet [26]	A feature pyramid network is utilized to extract multi-scale features, and cascade cost volumes are constructed to gradually narrow the depth range for coarse-to-fine depth prediction.	Low memory cost, adaptable to large-scale scenes, high-resolution and high-accuracy depth estimation and reconstruction is achieved.	Cascaded structure dependence on coarse predictions, potentially accumulating errors.	Regression
	UCSNet [27]	Based on variance uncertainty, adaptive layers are constructed to adjust depth planes and refine depth estimation.	Reconstruction resolution and accuracy are improved by flexible depth plane partitioning.	Variance-based strategy overlooks overall probability distribution, resulting in incomplete depth estimation.	Regression
	UniMVS [28]	Classification and regression are combined for deep prediction.	Cost volumes are directly constrained while sub-pixel accuracy is also achieved.	Heavily rely on classification, resulting in accumulating errors.	Unification

3) Extensive experiments on unmanned aerial vehicles (UAVs) and satellite remote sensing datasets have been conducted to compare the performance of the aforementioned SOTA networks with our CSC-MVS. The results indicate that the proposed network consistently outperforms other networks in remote sensing 3-D reconstruction tasks.

II. RELATED WORK

A. MVS Networks Based on Depth Map

The MVS network series revolves around the concept of plane sweeping [38]. As shown in Fig. 3, this method divides the given depth range into N_d parallel planes. Assuming a sufficiently dense distribution of planes, the true depth value of a point O on the surface of a spatial object must locate on one of the planes, D_i. Under ideal circumstances, point O should exhibit color consistency across different view images. In other words, if the correct depth value D_i is employed to project images from alternative views onto the reference view, there should be minimal or no difference at the corresponding pixel of point O. By evaluating N_d depths to construct pixel-level matching costs and selecting the plane depth value that achieves the best match for each pixel, the depth map of the scene can be obtained. Thus, the MVS networks design the following four modules to enable the end-to-end 3-D reconstruction.

1) Feature Extraction: To ensure the efficiency of the dense matching task, the MVS algorithm series first uses a 2-D convolutional neural network (CNN) structure to extract deep features. The input consists of N multiview RGB images $\{I_i\}_{i=1}^N$ (where $I_i \in \mathbb{R}^{H\times W\times C_I}$), which would be spatially downsampled in this module, resulting in N C_F-channel feature maps $\{F_i\}_{i=1}^N$, where $F_i \in \mathbb{R}^{H'\times W'\times C_F}$. In MVSNet [4], the downsampling reduces the spatial dimensions to $H' = H/4$ and $W' = W/4$. Subsequent improvements in the MVS network are designed from various perspectives, such as preserving spatial details by reducing downsampling operations [39], aggregating multiscale information through dense connections, expanding the spatial receptive field [24], or introducing attention mechanisms within views [25]. These improvements, however, often require the introduction of additional network modules or model parameters.

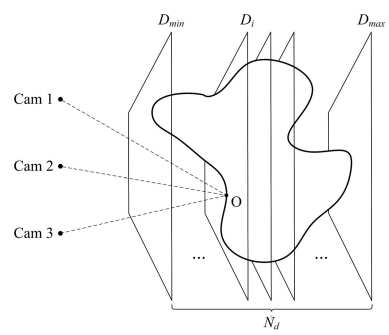


Fig. 3. Schematic illustrates the principle of plane-sweeping method.

2) 3-D cost Volume Construction: In the principle of plane sweeping [38], an important step involves performing geometric projection using the corresponding depth values to achieve

the warping of feature maps from the source view plane to the reference view plane. For datasets that provide camera intrinsic and extrinsic parameters, a differentiable homography transformation [40] is employed to achieve this process. The transformation can be represented as $x' \sim \boldsymbol{H}_i(d) \cdot x$, where "$\sim$" denotes the warping equation and $\boldsymbol{H}_i(d) \in \mathbb{R}^{3\times 3}$ represents the homography matrix between the ith source view feature maps and the reference view at depth d; however, for satellite images that deviate from the pinhole imaging model, an RPC warping based on a rational polynomial coefficients model will be used [16]. Similarly, this transformation can be expressed as $x' \sim \mathrm{RPC}^{\mathrm{fwd}}(h) \cdot \mathrm{RPC}^{\mathrm{inv}}(h) \cdot x$, where $\mathrm{RPC}^{\mathrm{fwd}}(h)$ and $\mathrm{RPC}^{\mathrm{inv}}(h)$, respectively, denote the forward and inverse warping processes between the world coordinate system points and the image coordinate system points at a ground height h (referred to as depth in the following text, conceptually similar to depth).

By employing the aforementioned transformations, the obtained feature maps $\{F_i\}_{i=1}^N$ are densely projected onto the reference view geometry based on N_d parallel depth planes, leading to the generation of feature volumes $\{V_i\}_{i=1}^N$ that are used to construct the cost volume, where $V_i \in \mathbb{R}^{H' \times W' \times N_d \times C_F}$. Then, to accommodate arbitrary numbers of input views, the model uses a variance-based cost metric to explicitly measure the matching similarity, thus aggregating the feature volumes into a cost volume C:

$$C = \frac{\sum_{i=1}^{N} \left(V_i - \bar{V}\right)^2}{N} \tag{1}$$

where $\bar{V}$ represents the average volume of the feature volumes.

MVSNet performs a single cost volume construction to obtain representations on the given N_d parallel depth planes, while subsequent research aims to optimize the cost volumes in a cascaded form [26], [41], [27]. This construction process typically relies on a feature pyramid with increasing scales. In the initial stage, low-resolution cost volumes are generated using low-scale features to estimate coarse depth maps. At this time, the depth planes cover the entire range uniformly, while as the resolution of the cost volumes increases, the subsequent depth range is gradually narrowed based on the predictions from the previous stage, resulting in depth maps of improved quality. Various methods can be used to refine the depth planes, including depth range sampling [26], depth estimation interpolation [41], and adaptive selection of intervals based on the variance uncertainty of the predictions [27]. These methods gradually partition the vast spatial extent of the scene by enhancing the accuracy and resolution of depth maps, enabling a complete depth reconstruction from coarse to fine. In terms of accuracy, considering the progressive relationship in the cascade, it is crucial to ensure the compatibility of the first-stage resolution with the scale of the scene.

3) Cost Volume Regularization: Next, the obtained cost volume is fed into a regularization module for optimization, aiming to mitigate initial noise from occlusions or nonsmooth surfaces. Initially, a multiscale 3-D UNet [42] network is employed in this module, using a four-level encoder–decoder structure to aggregate neighborhood information within a large receptive field. Upon the aggregation of the cost volume $C \in \mathbb{R}^{H' \times W' \times N_d \times C_F}$ at each level, a following single-layer convolution and softmax normalization along the depth dimension compress it into a single-channel probability volume $P \in \mathbb{R}^{H' \times W' \times N_d}$. This transformation allows the representation to be converted into the probability space of the depth planes, facilitating the subsequent depth prediction. Likewise, subsequent MVS networks have also introduced innovations in this section. For instance, recurrent regularization [5] treats the cost space as a concatenation of multiple depth planes and leverages recurrent neural networks [43], [44] to sequentially process the cost volume along the depth dimension. The sequential and recursive nature of this recurrent approach reduces memory consumption, rendering the model suitable for large-scale scene reconstruction.

4) Depth Estimation: To recover subpixel depth maps from the probability volume, most MVS networks use a technique called "soft argmin" in this module for prediction. Specifically, after obtaining the probability volume, the model calculates the expectation along the depth dimension, which corresponds to the weighted sum of the assumed depth values:

$$D = \sum_{d=d_{\min}}^{d_{\max}} d \times P(d) \tag{2}$$

where $P(d)$ represents the probability estimation at depth d. This computation enables the generation of continuous depth predictions, and the resulting depth map D has the same dimensions as the 2-D feature maps $\{F_i\}_{i=1}^N$. After resampling to the original size, the depth map is further optimized by applying the ℓ_1-norm loss on the depth-valid pixels.

In addition, certain models also employ the winner-take-all principle for depth prediction [5], [24], [25]; however, regardless of the specific computational approach, the optimization of depth estimation is consistently built upon the matching information. By relying solely on this constraint, the network must learn complex and extensive weight combinations, which introduces a potential risk of overfitting. In cases where the constraint originates from a single source, the difficulty of the model to resist noise and achieve accurate convergence, furthermore, increases, particularly in the context of remote sensing scenes with wider depth ranges and larger depth intervals.

It is evident that numerous studies have attempted to enhance the performance based on the seminal work of MVSNet [4], by modifying one or several modules among the four aspects mentioned above. The adjustment of the depth estimation module, in terms of the diversity of implementation methods and compatibility with different scenes, however, still requires further optimization. In this study, we propose CSC-MVS, an improved strategy for depth estimation through multitask optimization, aiming to enhance the model's adaptability to remote sensing scenarios.

B. Consistency Guidance

In recent years, several studies have explored the integration of consistency principles into multiview systems for assistance [45], [46]. Photometric consistency commonly assumes that corresponding pixels in different views share identical brightness and texture properties [32], [47], and similarly,

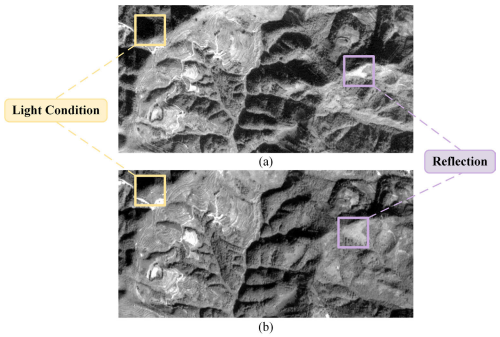


Fig. 4. Challenges that the photometric consistency assumption may encounter in the field of remote sensing. (a) Center view image. (b) Side view image.

cross-view consistency [48] and normal-depth consistency [49] focus on the pixels that have undergone cross-view projection or geometric transformations; however, the aforementioned methods generally suffer from the limitation that pixel-level matching assumptions encounter texture scarcity issues caused by practical factors such as illumination, noise, and occlusion, thereby failing to provide robust supervision signals [33]. This triggers significant challenges, especially for remote sensing data involving more complex scenes. For instance, as shown in Fig. 4, a comparison of images depicting exposed mountainous terrain from different views reveals cases where certain mountain folds do not align completely due to limitations, such as shadow occlusion or varying lighting conditions. Additionally, certain objects like mountaintop plateaus also exhibit differences in color and brightness among different views due to variations in reflection.

Confronting these challenges, it is natural to contemplate the idea of elevating pixel-level matching assumptions to a more abstract perceptual level, as the deep latent information embedded within images may overcome these limitations to some extent. Simultaneously, the abundant semantic information in remote sensing imagery provides a clear direction for exploring this approach. Based on this motivation, the concept of consistency can be extended from the pixel level to the feature level.

Inspired by the above insights, this study proposes CSC-MVS, which seeks breakthroughs in the depth estimation module by establishing consistency guidance at a higher-level semantic context. Subsequently, this guidance is seamlessly incorporated into the depth constraints through a multitask optimization objective, thereby facilitating comprehensive performance improvements in the MVS network.

III. Methods

CSC-MVS serves as a general improvement strategy applicable to the majority of current deep learning-based MVS algorithms, enabling seamless integration with any MVS network by introducing additional branches. To strike a balance between the reliability of semantic consistency and data dependency, we establish branches in both global and local aspects and adopt appropriate clustering methods for each. These branches are connected after a pretrained VGG feature extractor [50] to exploit underlying semantics; furthermore, when incorporating the extracted semantics into the optimization objective, we adopt an uncertainty multitask loss [51] to ensure the plug-and-play and lightweight nature of our method. The overall architecture of CSC-MVS is illustrated in Fig. 5, and the individual branches will be discussed in detail in this section.

A. Depth Estimation Network Backbone

In theory, CSC-MVS can be combined with any network in the MVS series, and here, we take the classic model MVSNet [4] as a representative example. The pipeline diagram is shown at the bottom of Fig. 5, where a single reference view image and $N-1$ source view images are fed into the network. The MVS backbone network then performs feature extraction and uses geometrical parameters to construct a matching cost volume, which is subsequently regularized to estimate an initial depth map of the reference view image.

B. Nonnegative Matrix Factorization (NMF) Local Semantic-Supervised Branch

NMF [52] has gained considerable attention for its interpretability in analyzing matrix data [53]. Recently, it has been observed that combining NMF with CNNs can capture local semantic correspondences in the feature space [54]. Inspired by this, we introduce the NMF algorithm to acquire semantics from different views and propose the local semantic-supervised branch that exploits the cross-view consistency to assist in optimizing depth estimation. In this section, we begin by explaining the fundamental principles of NMF.

According to the theory of NMF [52], any nonnegative matrix $\mathbf{A} \in \mathbb{R}_+^{m\times n}$ can be decomposed into the product of two nonnegative matrices $\mathbf{B} \in \mathbb{R}_+^{m\times K_L}$ and $\mathbf{Q} \in \mathbb{R}_+^{K_L\times n}$, such that $\mathbf{A} \approx \mathbf{BQ}$. The matrix $\mathbf{B}$ is known as the basis matrix, while the matrix $\mathbf{Q}$ is referred to as the coefficient matrix. In the case where the matrix $\mathbf{Q}$ is subject to the orthonormal constraint $\mathbf{QQ}^{\mathrm{T}} = \mathbf{I}$, the Frobenius norm can be achieved by minimizing the following error function [55]:

$$|\mathbf{A} - \mathbf{BQ}|_F^2, \text{ s. t. } \mathbf{B}, \mathbf{Q} \geq 0. \tag{3}$$

It is apparent that the decomposed matrix $\mathbf{B}$ can be effectively used to approximate matrix $\mathbf{A}$. The dimension K_L of $\mathbf{B}$, defined as the predefined rank in this approximation process, can also be interpreted as the number of semantic clusters from a clustering perspective.

Under the assumptions of nonnegativity and orthonormal constraints, NMF provides an approximation of the original matrix $\mathbf{A}$ as a weighted mapping constructed from K_L basis vectors, which are the column vectors of the basis matrix $\mathbf{B}$ [53]. In this context, the matrix $\mathbf{B}$ contains the semantic information of K_L clustering objects, and they are linearly independent of each other [56]. This additive combination representation based on the basis vectors possesses a more intuitive semantic meaning. Its fundamental idea of "local components forming the whole" distinguishes it from partial clustering methods that solely obtain image representations. Meanwhile, this approach enables the exploration of local

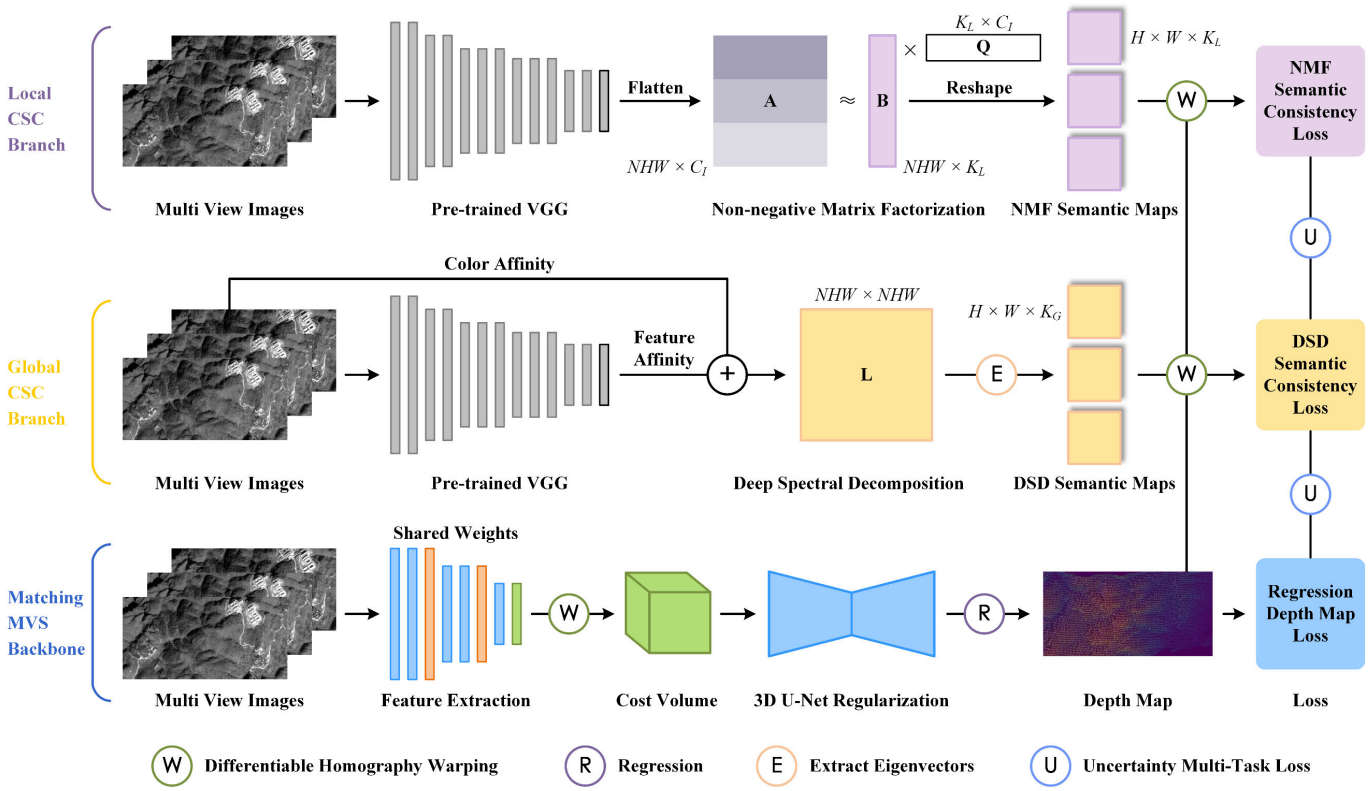


Fig. 5. Overall architecture of the proposed CSC-MVS method consists of three components arranged from top to bottom: the NMF local branch (see Fig. 6), the DSD global branch (see Fig. 7), and the MVS network backbone. Through the differentiable homography warping on the right, the information in the semantic maps will be integrated with the depth map. Then, these components are connected by the final uncertainty multitask loss module.

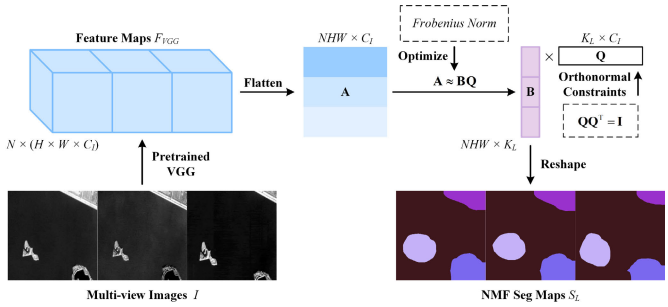


Fig. 6. Structure of the NMF local semantic-supervised branch. The italic letters next to the feature maps or matrices indicate their respective shapes. The different colors in the semantic maps SL represent different semantic classes.

semantic objects corresponding across different views while minimizing the optimization cost. As illustrated in the obtained semantic maps in Fig. 6, the NMF algorithm demonstrates the ability to recognize and differentiate local entities in the original image, capturing the positional variations of objects across multiple views.

Building upon the NMF decomposition of the original images, its feature space enables the discovery of deeper semantic features [54]. Considering the high data costs and complex application scenarios in the remote sensing field for MVS tasks, we combine the selected feature extractor with the NMF algorithm. This results in the formation of our local semantic-supervised module, which can capture deep semantic similarities across multiple views in an unsupervised way and thus generate constraints for optimizing depth estimation through cross-view warping. Based on this, we establish the architecture of this branch. As shown in Fig. 6, we first apply a pretrained VGG network [50] to perform feature extraction on the multiview images $\{I_i\}_{i=1}^{N}$, where $I_i \in \mathbb{R}^{H\times W\times C_I}$. Subsequently, to extract pixel-level semantic information, the obtained feature maps are connected and flattened and then subjected to NMF in the form of an $NHW \times C_I$ matrix $\mathbf{A}$, resulting in an $NHW \times K_L$ matrix $\mathbf{B}$. By reshaping the matrix $\mathbf{B}$, we obtain N local semantic maps S_L of size $H \times W \times K_L$ corresponding to the multiple views, which are used for cross-view consistency guidance.

C. Deep Spectral Decomposition (DSD) Global Semantic-Supervised Branch

To further enhance CSC-MVS, we incorporate a global clustering-based method that complies with cross-view correspondences. Research on graph theory has demonstrated that the global properties of a graph are closely related to the eigenvalues and eigenvectors of its Laplacian matrix [57], [58], [59]. Leveraging this correlation, previous studies have used spectral decomposition algorithms to assist unsupervised semantic localization and segmentation [60]. Motivated by these findings, we introduce the DSD algorithm positioned after the VGG feature extractor, thus proposing a module that enables the extraction of multiview global semantics and leverages their consistency to form constraints for depth estimation.

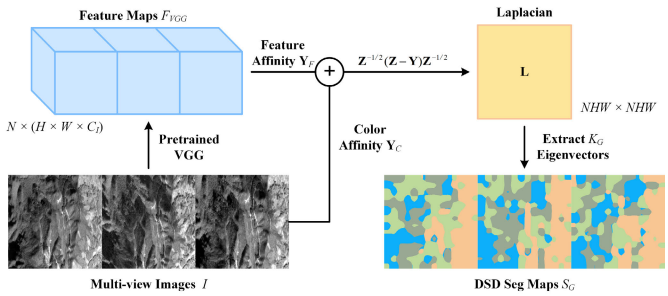


Fig. 7. Structure of the DSD global semantic-supervised branch. The italic letters next to the feature maps or matrices indicate their respective shapes. The symbol "+" in the middle refers to element-wise addition after weighting the two matrices. The different colors in the semantic maps SG represent different semantic classes.

In graph theory [57], a weighted undirected graph $G = (V, E)$ is first used to represent the input image for segmentation, where the vertices V correspond to the image pixels and the edges E capture the connections between adjacent pixels v_i and v_j. The weight $y(v_i v_j)$ that assigned to each edge quantifies the neighborhood similarity of pixels in terms of grayscale, color, or texture, forming the G's adjacency matrix $\mathbf{Y} = \{y(v_i v_j) : v_i v_j \in E\}$. The weighted degree matrix $\mathbf{Z}$ is obtained by summing the rows of the adjacency matrix $\mathbf{Y}$, with the diagonal elements containing the sum of edge weights associated with each vertex (i.e., each pixel). Then, the normalized Laplacian matrix of the graph is defined as follows:

$$\mathbf{L} = \mathbf{Z}^{-1/2}(\mathbf{Z} - \mathbf{Y})\mathbf{Z}^{-1/2}. \tag{4}$$

By performing the spectral decomposition on the obtained Laplacian matrix, the crucial eigenvalues and eigenvectors can be derived.

The principle of graph cut dictates that the optimal segmentation of an image should follow the criterion of maximizing the similarity within each segmented region while minimizing the similarity between different regions [61]. Meanwhile, the previous works of Donath and Hoffman [62] and Fiedler [63] have confirmed that the eigenvectors of the graph Laplacian operator yield the globally minimal energy graph partitioning. As a result, the semantic map based on graph cut is no longer characterized by independent and discrete blocks but rather exhibits a more comprehensive representation of the image content (see Fig. 7).

Based on the principles described above, we have devised the DSD global semantic-supervised branch, as illustrated in Fig. 7. First, the multiview feature maps denoted as F_{VGG} are extracted using a pretrained VGG network [50]. Next, it is necessary to construct an appropriate $\mathbf{Y}$ matrix. To fully use multilevel information, we opt to aggregate the color-space adjacency matrix $\mathbf{Y}_C$ derived from the multiview images, with the feature adjacency matrix $\mathbf{Y}_F$ computed from the obtained feature maps (i.e., $\mathbf{Y} = \mathbf{Y}_C + \mathbf{Y}_F$).

In the first step, we integrate the color information and spatial location attributes of the original image using the K-nearest neighbors (KNNs) algorithm [64]. For each pixel i in the multiview images, its feature vector in the HSV color space is defined as follows:

$$X(i) = (\cos(h), \sin(h), s, v, x, y). \tag{5}$$

Here, (h, s, v) and (x, y) correspond to the HSV coordinates and spatial coordinates of pixel i, respectively. Let j be the neighboring pixel of i, the color-space adjacency matrix $\mathbf{Y}_C$ of the original image can be obtained by considering the pixel-level KNN of the feature vectors:

$$\mathbf{Y}_C = 1 - \|X(i) - X(j)\|, \quad i \in KNN(j). \tag{6}$$

Subsequently, following the approach in [60], we aggregate the positive self-correlations of F_{VGG} (i.e., $F_{\text{VGG}} F_{\text{VGG}}^{\text{T}} > 0$) and calculate the feature adjacency matrix $\mathbf{Y}_F$ using the inner product:

$$\mathbf{Y}_F = F_{VGG} F_{VGG}^{\text{T}} \odot \left(F_{VGG} F_{VGG}^{\text{T}} > 0\right). \tag{7}$$

Finally, the adjacency matrix $\mathbf{Y}$ used for computing the Laplacian matrix and performing spectral decomposition is obtained by weighted adding the above two matrices.

After obtaining the adjacency matrix $\mathbf{Y}$, we proceed to compute the multilevel Laplacian matrix $\mathbf{L}$ that integrates low-level color and deep features using (4). By applying spectral decomposition on matrix $\mathbf{L}$, the eigenvectors corresponding to the top K_G positive real-valued eigenvalues are extracted. Similar to S_L, these eigenvectors are reshaped to obtain the multiview global semantic maps $\{\mathbf{S}_G^i\}_{i=1}^N$, which can also be used for consistency guidance and $\mathbf{S}_G^i \in \mathrm{R}^{H \times W \times K_G}$.

D. Depth Estimation Optimization and Uncertainty Multitask Method

Different from the direct supervision of depth by the MVS framework, the semantic-supervised branch assists depth optimization from the perspective of view consistency. When the depth values estimated by the backbone network are sufficiently accurate, the semantic features after projection from source views, should be consistent with that of the reference view. This implies that the cross-view consistency of the semantic map is positively correlated with the accuracy of the depth map. Therefore, we propose two semantic-supervised modules specifically designed for the remote sensing field to establish cross-view consistency in both global and local directions, achieving multiple optimizations for depth estimation. The specific optimization objectives include: 1) depth estimation constraint based on matching information and 2) global and local semantic consistency constraints. After constructing these two optimization tasks, we further design: 3) an uncertainty-based multitask weighting method [51] to organically combine the semantic consistency loss with the depth estimation loss. The details are presented below.

1) Depth Estimation Constraint: For the depth estimation loss, we take the case of MVSNet [4] as the backbone network and provide its computation formula as follows:

$$L_{DE} = \sum_{p \in P_{\text{vali,}}} \|d(p) - \hat{d}(p)\|_1. \tag{8}$$

Here, p refers to the depth-valid pixel in the reference view, $d(p)$ and $\hat{d}(p)$ represent the ground truth value and the estimated value of the corresponding depth, respectively.

2) Global and Local Semantic Consistency Constraints: Based on the depth estimation, given the predicted depth value $\hat{d}(p)$, we consider p' as the pixel in source view i corresponding to p. The semantic map $S_i^{src}(p')$ on p' can be projected to the corresponding pixel p in the reference view through homography warping or RPC warping, which can be expressed as:

$$S_i^{\text{warp}}(p) \sim S_i^{src}(p'). \tag{9}$$

Subsequently, the semantic consistency discrepancy is measured by the cross-entropy loss between the reference view's semantic map $S_1^{\text{ref}}(p)$ and the warped $S_i^{\text{warp}}(p)$ on depth-valid pixels. For semantic consistency constraints, the calculation formulas for the global loss L_{SC}^G and local loss L_{SC}^L can be summarized as:

$$L_{\text{SC}}^j = \sum_{i=2}^{N} \times \left[\frac{1}{\|M_i\|_1} \left(-\sum_{p \in M_i} f\left(S_{1,j}^{\text{ref}}(p)\right) \log\left(S_{i,j}^{\text{warp}}(p)\right) M_i \right) \right] \quad j \in \{G, L\} \tag{10}$$

where M_i represents the mask for valid pixels, and the function $f(\cdot)$ denotes one-hot encoding for $S_1^{\text{ref}}(p)$.

3) Uncertainty Multitask Method: Through the above calculation formulas, we obtain three reliable losses: L_{DE}, L_{SC}^G, and L_{SC}^L; however, a key issue arises: how to appropriately measure the relative influences of the MVS backbone, global semantic branch, and local semantic branch on depth optimization. Manual tuning to discover the optimal weight distribution is time-consuming and laborious. To address this issue and enable flexible allocation of the optimization functions in a multitask form, this study introduces the use of an uncertainty parameter σ to achieve adaptive assignment of the weights for the three losses.

Specifically, we propose a weighted loss function based on task uncertainty, enabling the convenient and efficient dynamic update of the overall network parameters. The formula is defined as:

$$\text{Loss} = \sum_{i=1}^{3} (w_i L_i + r_i). \tag{11}$$

Here, w_i and r_i represent the weight and regularization term of the corresponding loss L_i, respectively. Recognizing that the matching information remains central to depth estimation and optimization, we fix the weight of the depth estimation loss to 1. The relative weights of the other two semantic consistency losses are learned through network training. That is, for the depth estimation loss L_{DE}:

$$w_{L_{\text{DE}}} = 1.0, r_{L_{DE}} = 0.0. \tag{12}$$

For the semantic consistency loss L_{SC}, we have:

$$w_{L_{\text{SC}}^j} = \exp\left(-\log \sigma_j^2\right), r_{L_{\text{SC}}^j} = \log \sigma_j^2, j \in \{G, L\} \tag{13}$$

where σ_j represents the uncertainty of the task, and the associated value $\log \sigma_j^2$ is adaptively adjusted through network parameter learning. Finally, the loss applied in our method CSC-MVS, denoted as Loss, is obtained by summing the three losses according to the defined weights and regularization terms in (11).

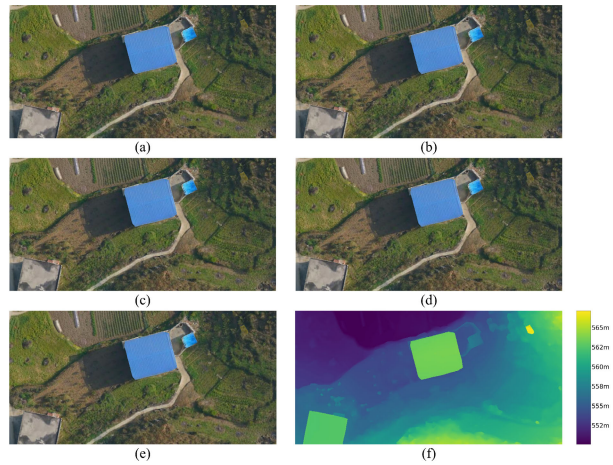


Fig. 8. Sample for the LuoJia-MVS dataset consists of the images: (a) left-side view, (b) right-side view, (c) bottom-side view, (d) top-side view, (e) center view, and (f) ground truth depth map.

IV. Results

A. Datasets

For the evaluation of our proposed method, we employed two datasets: the LuoJia-MVS dataset [15], constructed from aerial imagery, and the WHU TLC dataset [16], composed of optical satellite images.

1) LuoJia-MVS Dataset: This dataset was constructed by projecting a 3-D surface model derived from thousands of stereo aerial images. It comprises a total of 5680 groups of five-view images, accompanied by pixel-level depth maps and accurate camera parameters. Each image in the dataset has a size of 784 × 368 and a spatial resolution of 10 cm. The dataset was divided into training and testing sets in a ratio of approximately 3:1, with 4320 groups of images used for training and 1360 groups for testing [15]. Examples of the five-view images and corresponding ground truth depth maps can be seen in Fig. 8.

2) WHU TLC dataset: The three-view images in this dataset were acquired by the Ziyuan-3 (ZY-3) satellite, which is equipped with TLC cameras capable of simultaneous imaging. The spatial resolutions of the bottom nadir image and the two side oblique images are 2.1 and 2.5 m, respectively. After preprocessing, these images were partitioned into 6802 groups of 768 × 384 size image patches. Among them, 5011 were used for training, while the remaining groups were used for testing [16].

Unlike the LuoJia-MVS dataset, the ground truth provided in this dataset is derived from a digital surface model (DSM) with a ground resolution of 5 m, which is generated through laser scanning and ground control points [16]. This DSM is then projected onto the reference image, providing height information for the corresponding pixels in the satellite images. As discussed in Section II of this article, the height

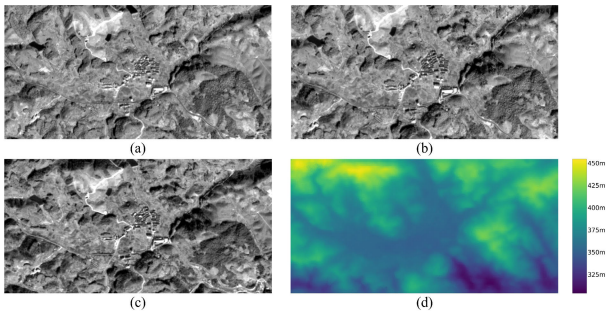


Fig. 9. Sample for the WHU TLC dataset consists of the images: (a) left-side view, (b) right-side view, (c) center view, and (d) ground truth depth map.

value h plays a role like the depth value d in close-range MVS, serving as an intermediary during the cross-view warping process. To ensure subpixel accuracy in the projection, the dataset includes precalibrated RPC parameters. Fig. 9 illustrates examples of the three-view images and the corresponding ground truth depth map in this dataset.

B. Implementation

All the experiments were conducted on a desktop computer using PyTorch 1.1.0 with an Intel Xeon Silver 4210 CPU (2.20 GHz), 128 GB RAM, and an 11 GB GeForce RTX 2080Ti GPU.

For the seven comparative methods, the number of epochs for network training was set to 35, and the RMSprop optimizer with a momentum parameter $\alpha = 0.9$ was adopted. The initial learning rate for UCSNet was set to 0.0016, according to [27], while for the other networks, it was set to 0.001. After 10 epochs, the learning rate was reduced by a factor of 2 every two epochs. For the experiments on the LuoJia-MVS dataset, as suggested in [15], a batch size of one unit was used. For the experiments on the WHU TLC dataset, due to the computational memory and time required by RPC warping, two GPUs were used. Consequently, the batch size for these experiments was set to 2.

Referring to the settings in the original paper [16], all the experiments conducted on the WHU TLC dataset followed specific parameter configurations. The number of height hypotheses was initially fixed at 64, and the height interval was determined based on this value and the provided image height range in the RPC parameters. For the cost volume in noncascade networks, the dimension during construction was set to 32, and the downsampling ratio relative to the original multiview images was set to 1/16 to conserve memory resources. In the case of cascaded networks, parameters needed to be specified for each of the three stages. The number of depth hypotheses planes in the third stage was successively set to {64, 32, 8}, and the dimensions and relative downsampling ratios of the cost volumes were set to {32, 16, 8} and {1/16, 1/4, 1}, respectively. Since each stage has a different depth resolution, their weights for loss were set to {0.5, 1.0, 2.0} in order. Except for UCSNet, which implemented its own adaptive interval strategy, the height interval values for the three stages of the other networks were set to $\{(d_{\max} - d_{\min})/64, 5\text{ m}, 2.5\text{ m}\}$.

Regarding the experiments on the LuoJia-MVS dataset, we followed the parameter settings described in the existing methods [6], [8]. For the noncascade networks, the depth hypotheses and depth intervals were set to 192 and 1, respectively. The dimensions of the cost volume and the downsampling ratio were set to 32 and 1/16, respectively. For the cascade networks, we adopted the parameter settings from the relevant literature [15]. Specifically, the depth hypotheses and depth intervals were {48, 32, 8} and {4, 2, 1} from the first stage to the third stage, respectively. The dimensions of the cost volume and the downsampling ratio were {32, 16, 8} and {1/16, 1/4, 1}, respectively. The weights for loss at each stage remained {0.5, 1.0, 2.0}. The special handling for UCSNet, furthermore, remained the same as described above.

C. Other Configurations

Some parameters affecting the convergence of the proposed network are also associated with the NMF and DSD branches. First, to maintain consistent semantic representation, both branches share the same dimension for decomposition. Within the recommended range [51], we ultimately selected $K_s = K_G = K_L = 4$, which is determined through sensitivity analysis (see Section V-D). NMF clustering was initialized with a random seed, and the maximum number of iterations and tolerance factor were set to 50 and 1e-4, respectively [51]. Considering that failure to converge renders the decomposition component (i.e., with NaN values) unsuitable for back-propagation, we reinitialized the parameters and conducted NMF decomposition again until convergence. The parameter of DSD clustering used to balance the color-space and feature matrix was set to 10.0 [57]. Regarding the initial uncertainty parameter for the weight of these two branch losses, it was set to $\sigma_G = \sigma_L = 1.0$ to ensure adequate fitting performance in terms of dense matching.

D. Accuracy Assessment

The main evaluation metric used in this article is the mean absolute error (MAE) [6], [65], [66], which measures the average ℓ_1-norm difference between the estimated depth values and the ground truth depth values; moreover, considering the characteristics of the two datasets, we defined appropriate threshold metrics for each to assess the completeness of the methods. For the WHU TLC dataset, we used the <7.5 m (%) and <2.5 m (%) metrics [16], while for the LuoJia-MVS dataset, we adopted the <0.6 m (%) and <3-interval (%) metrics [15]. These threshold metrics indicate the percentage of pixels with prediction error within threshold requirements relative to the total number of valid pixels. The term "3-interval" represents three times the corresponding depth interval value. Considering the 10 cm resolution of the LuoJia-MVS dataset, this value is expected to fluctuate around 0.3 m.

E. Performance Evaluation

The quantitative comparative results involving the seven MVS baseline networks demonstrate that the proposed

TABLE II

DEPTH MAP RECONSTRUCTION ACCURACY OF THE SEVEN SOTA NETWORKS BEFORE AND AFTER APPLYING THE CSC-MVS METHOD ON THE WHU TLC DATASET

Model Type	Method	MAE (m)	<2.5m (%)	<7.5m (%)
Non-Cascade	MVSNet	2.30	63.5	93.8
	MVSNet + CSC-MVS	**2.15**	**66.8**	**94.7**
	RMVSNet	2.23	63.9	95.0
	RMVSNet + CSC-MVS	**2.18**	**65.6**	**95.2**
	D2HC-RMVSNet	2.21	65.6	93.9
	D2HC-RMVSNet + CSC-MVS	**2.09**	**68.2**	**95.0**
	AA-RMVSNet	2.35	62.6	93.8
	AA-RMVSNet + CSC-MVS	**2.32**	**63.0**	**94.0**
Cascade	Cas-MVSNet	2.19	66.3	93.6
	Cas-MVSNet + CSC-MVS	**2.04**	**69.2**	**94.9**
	UCSNet	2.34	62.7	92.7
	UCSNet + CSC-MVS	**2.30**	62.7	**93.7**
	UniMVS	2.67	55.2	**91.7**
	UniMVS + CSC-MVS	**2.60**	**56.8**	91.2

TABLE III

DEPTH MAP RECONSTRUCTION ACCURACY OF THE SEVEN SOTA NETWORKS BEFORE AND AFTER APPLYING THE CSC-MVS METHOD ON THE LUOJIA-MVS DATASET

Model Type	Method	MAE (cm)	<0.6m (%)	<3-intv. (%)
Non-Cascade	MVSNet	17.4	96.1	92.4
	MVSNet + CSC-MVS	**17.1**	96.1	**92.5**
	RMVSNet	17.7	96.0	**93.5**
	RMVSNet + CSC-MVS	**16.4**	**96.4**	93.1
	D2HC-RMVSNet	16.9	96.2	92.6
	D2HC-RMVSNet + CSC-MVS	**16.1**	**96.5**	**93.3**
	AA-RMVSNet	31.0	89.7	88.9
	AA-RMVSNet + CSC-MVS	**29.5**	**90.9**	**90.4**
Cascade	Cas-MVSNet	10.3	98.4	97.1
	Cas-MVSNet + CSC-MVS	**9.5**	**98.6**	**97.6**
	UCSNet	10.7	98.5	97.4
	UCSNet + CSC-MVS	**10.0**	98.5	**97.5**
	UniMVS	14.1	97.6	95.7
	UniMVS + CSC-MVS	**13.2**	**97.8**	**96.3**

CSC-MVS method achieves effective performance improvement over existing SOTA methods for both aerial image-based and optical satellite image-based 3-D reconstruction tasks. As indicated in Tables II and III, when considering the MAE metric alone, CSC-MVS shows remarkable positive gains across all the SOTA networks. In the case of classical MVS networks, such as MVSNet and Cas-MVSNet, the CSC-MVS method achieves substantial improvements in the range of 5%–8%. Meanwhile, for other baselines like R-MVSNet and D2HC-RMVSNet, the CSC-MVS method also yields accuracy gains ranging from 2% to 5%. In terms of completeness, the CSC-MVS method, moreover provides a certain degree of assistance in enhancing the completeness of the depth map reconstruction for most cases, although it slightly exhibits instability with regard to its accuracy gains. Nevertheless, overall, the trends in both accuracy and completeness metrics closely align, further underscoring the comprehensive advantages of proposed CSC-MVS algorithm in improving reconstruction outcomes.

The effectiveness and generality of the proposed CSC-MVS method can be validated through the experimental results on both datasets. These results, however, also raise some intriguing questions for further consideration. First, one notable observation is that the overall accuracy and completeness of the WHU TLC dataset are relatively lower compared to the LuoJia-MVS dataset. We attribute this difference to several characteristics of satellite remote sensing data, including the lower spatial resolution, the wide coverage of scenes, and the complexity of the Earth's surface landscapes [67]. ZY-3, as China's first professional satellite dedicated to high-resolution stereo mapping, enables large-scale reconstruction of the Earth's surface [68], [69], and deep learning-based MVS methods can effectively serve this purpose [70]. Consequently, the experimental results on the WHU TLC dataset are persuasive and can further demonstrate the applicability of our proposed method in the remote sensing field.

Second, the gains obtained through the multistage cascade architecture exhibit noticeable discrepancies between the two datasets. As observed from the comparisons in the tables, the cascade design improves the accuracy on the LuoJia-MVS dataset by approximately 17%–42%, whereas on the WHU TLC dataset, the improvement ranges from only 1%–7%. This phenomenon suggests that the cascade structure is more beneficial for object reconstruction in aerial imagery, while the coarser spatial resolution of the imagery and the larger scale variations of the object in the satellite scene lead to smaller gains in the depth estimation task for the method. Notably, the performance improvement achieved by CSC-MVS is consistent for both cascade and noncascade networks. In fact, some noncascade networks even outperform the cascade networks when combined with our approach. This implies the robustness of our method in handling diverse remote sensing scenarios.

The qualitative visualization results of depth maps further highlight the performance enhancements of our method over SOTA networks in various application scenarios. Following the analysis above, we selected two representative networks from each group, namely MVSNet, D2HC-RMVSNet, Cas-MVSNet, and UCSNet. Several examples for visual comparison of these four methods are presented in Fig. 10. In the WHU TLC dataset, which primarily covers mountainous and surrounding areas, we chose typical examples of bare mountains with scattered houses in Fig. 10(a) and dense construction sites in Fig. 10(b) for comparison. Additionally, to ensure the diversity of the examples, we selected vegetation-lush forest areas in Fig. 10(c) and rural residential areas with cultivated land in Fig. 10(d) from the LuoJia-MVS dataset, which has a more varied land cover. The visual comparisons will be more obvious in the residual maps, which consist of relative residuals calculated by dividing the absolute difference between the reference depth and inferred depth by the reference depth.

In general, for regions with homogeneous surfaces, such as open fields and cultivated land, the residual values tend to be low. In contrast, heterogeneous objects, like forests and bare mountains, occasionally show low-value residual artifacts; furthermore, higher residual values are commonly distributed along the edges of ground objects. Upon analyzing the results of the baselines after applying the CSC-MVS method in Fig. 10, it is apparent that this approach can mitigate larger residuals at the object boundaries. At the same time, the

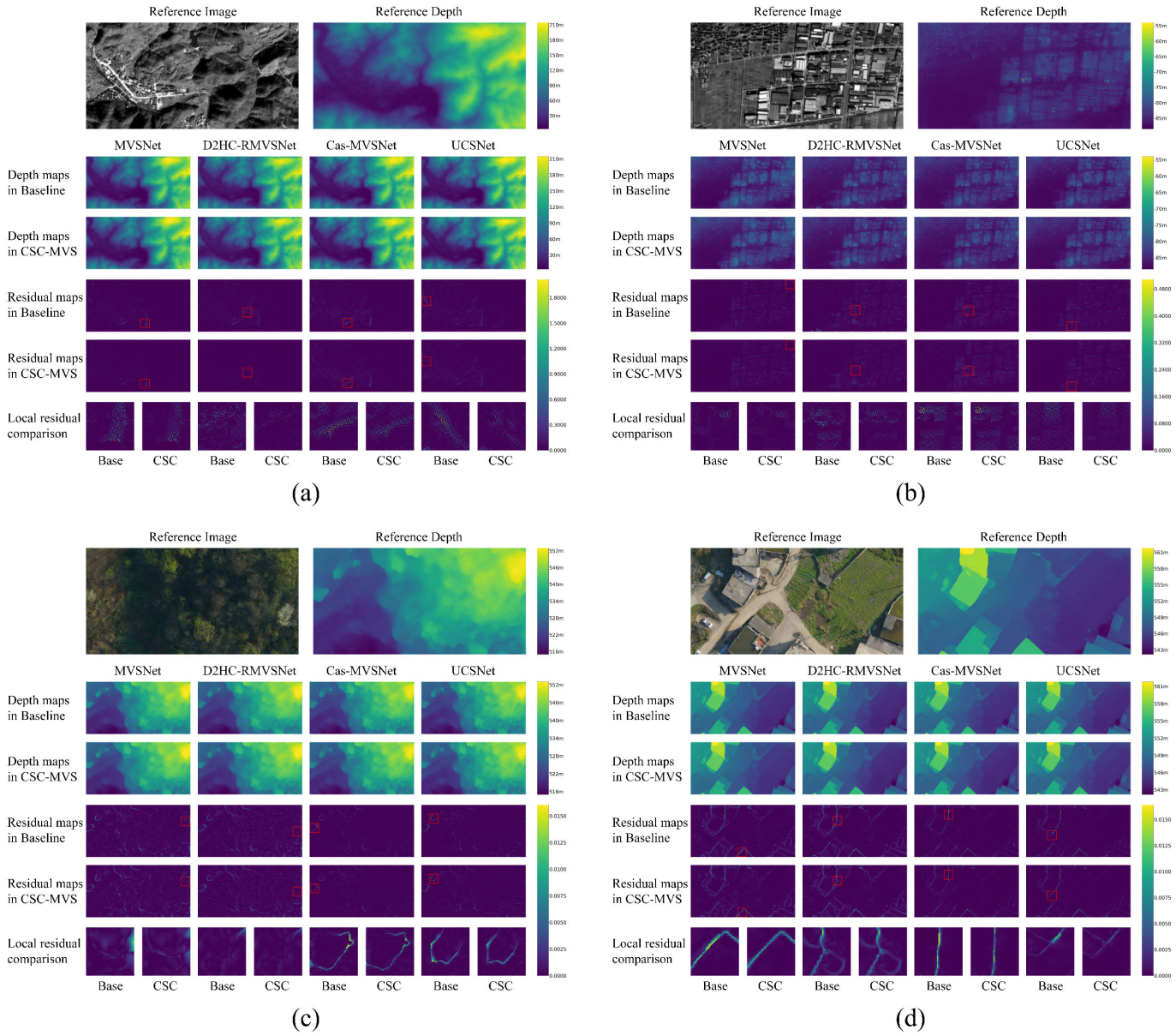


Fig. 10. Visual comparison of several SOTA networks before and after applying the CSC-MVS method. (a) Mountain area with scattered houses. (b) Construction site with dense factory buildings. (c) Verdant forest land. (d) Rural residential area with cultivated land. Each sub-figure is organized as follows. The first row shows the reference image (left) and the ground truth depth map (right). The second and third rows display the inferred depth maps of MVSNet, D2HC-RMVSNet, Cas-MVSNet, and UCSNet, both in the original case (top) and the improved case (bottom) by applying the CSC-MVS method. The fourth and fifth rows represent the relative residual maps of the baseline networks (top) and the improved networks (bottom). The last row presents a local magnification of the noticeable improvement in the residuals (highlighted by red boxes in the previous two rows). Each model showcases a comparison between the state before (left, denoted as "Base") and after improvement (right, denoted as "CSC").

residual artifacts that originally existed in Fig. 10(b) and (c) were also effectively alleviated through the method.

The observed improvements are majorly attributed to the proposed semantic consistency branches, which possess dual properties of attribute and spatial location. In this context, the local semantic consistency generated by the NMF branch strengthens the supervision signal on individual objects, leading to refined depth estimation for these objects and their boundaries. Meanwhile, the DSD branch emphasizes and supplements global information, effectively compensating for estimation biases that tend to exhibit agglomerate distributions. In summary, as a deep and intrinsic characteristic of images, semantic information remains robust despite challenges, such as coarse ground resolution and steep depth intervals; thus, it can effectively characterize complex scenes and demonstrate good adaptability to various remote sensing scenarios. This auxiliary approach, moreover, retains the advantages of the baseline networks while minimizing the potential loss of original matching information, thereby ensuring compatibility with multiple MVS methods.

V. Discussions

A. Consistency Comparison

In this section, we evaluate the effectiveness of using feature-level semantics for consistency guidance by comparing

TABLE IV
PERFORMANCE COMPARISON OF SEVERAL CONSISTENCY METHODS ON THE WHU TLC DATASET

Method	MAE (m)	<2.5m (%)	<7.5m (%)
Baseline (MVSNet)	2.30	63.5	93.8
+ Photometric Consistency [47]	2.27	64.3	93.6
+ Normal-depth Consistency [49]	2.26	64.3	93.5
+ Photometric + Normal-depth Consistency	2.29	63.6	93.7
+ Semantic Consistency (UNet) [42]	2.19	65.6	94.4
+ Semantic Consistency (Proposed)	**2.15**	**66.8**	**94.7**

our proposed semantic consistency method with two alternative methods: photometric consistency [47] and normal-depth consistency [49]. The former measures the photometric consistency using an ℓ_1-norm loss and incorporates additional guidance signals, such as image gradients, structured similarity, and depth smoothness to enhance robustness. The latter further introduces the normal-depth consistency term based on the orthogonality between normals and local surface tangents. It uses the geometric transformation relationship and the regularization process to jointly refine depth in both 2-D and 3-D spaces. Considering the credibility of the evaluation, we conduct experiments on the more challenging satellite benchmark, the WHU TLC dataset [16]. To ensure fairness, we use MVSNet as the baseline for comparison. In addition, we also compared the clustering approach with the classic semantic segmentation method UNet [42] trained from the WHDLD dataset [71], which is similar to the WHU TLC and is derived from high-resolution imagery captured by the Gaofen-1 and Ziyuan-3 satellites, with a resolution of 2 m and containing six classes, including buildings, vegetation, bare soil, and others. The specific experimental results are summarized in the table below, with the best metrics highlighted in bold.

Table IV illustrates the performance of MVS networks with various consistency enhancement methods. In general, the introduction of any form of consistency guidance has a positive impact on the baseline performance; however, it can be observed from the table that our proposed semantic consistency method yields the best results. First, our clustering method without additional semantic annotations is superior to the classic semantic segmentation method UNet. Second, compared to the photometric consistency and the normal-depth consistency, our proposed semantic consistency demonstrates significant improvements in both accuracy and completeness. Interestingly, the combination of photometric consistency and normal-depth consistency shows a weakened enhancement effect. This finding further suggests that the pixel-level guidance signals of the two methods are relatively weak and less applicable in complex remote sensing scenarios, thus highlighting the validity of our proposed approach.

B. Ablation Study

To further investigate the rationality and necessity of the designed semantic consistency guidance branch and uncertainty multitask optimization method, we conducted ablation experiments on the WHU TLC dataset using the MVSNet and Cas-MVSNet baselines in both noncascade and cascade configurations.

TABLE V
ABLATION STUDY RESULTS FOR BOTH THE NMF BRANCH AND THE DSD BRANCH IN THE PROPOSED CSC-MVS METHOD ON THE WHU TLC DATASET

Model	Semantic Consistency Module	MAE (m)	<2.5m (%)	<7.5m (%)
MVSNet (Non-Cascade)	Baseline	2.30	63.5	93.8
	+ NMF	2.20	65.5	93.8
	+ DSD	2.29	63.6	93.9
	+ Both	**2.15**	**66.8**	**94.7**
Cas-MVSNet (Cascade)	Baseline	2.19	66.3	93.6
	+ NMF	2.16	66.8	94.3
	+ DSD	2.10	68.3	94.4
	+ Both	**2.04**	**69.2**	**94.9**

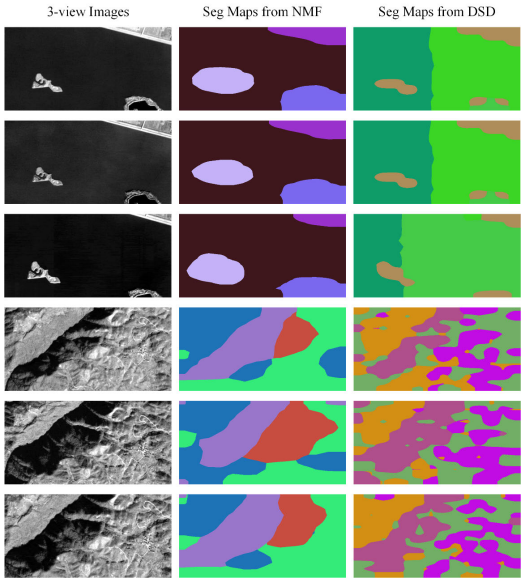


Fig. 11. Original three-view images (left) and the visual semantic maps obtained through the NMF branch (middle) and through the DSD branch (right).

1) Semantic Consistency Guidance Branch: Table V is divided into two parts based on whether cascade is used or not. Each part begins with the baseline network, followed by versions with the NMF branch, DSD branch, and both branches combined. The results in the table demonstrate that for both noncascade and cascade networks, using either the NMF branch or the DSD branch individually leads to a 1%–4% improvement in accuracy compared to the baseline. Notably, combining both branches lead to a significant improvement of 7%. This indicates that the two proposed branches in this article complement each other in enhancing the performance of MVS networks in depth map reconstruction tasks.

To delve deeper into the role of the two semantic consistency branches proposed in this article, we present visual results of several examples from the WHU TLC dataset in Fig. 11, showcasing the effects of each branch individually. The visual analysis reveals that both methods attach importance to semantic extraction and exhibit the ability to locate prominent objects. Particularly in the NMF semantic maps shown in the middle column, the lake platforms and mountain

TABLE VI

ABLATION STUDY RESULTS FOR THE UNCERTAINTY MULTITASK OPTIMIZATION IN THE PROPOSED CSC-MVS METHOD ON THE WHU TLC DATASET

Model	Seg Weight Setting	MAE (m)	<2.5m (%)	<7.5m (%)
MVSNet (Non-Cascade)	W_NMF = 0, W_DSD = 0	2.30	63.5	93.8
	W_NMF = 0.01, W_DSD = 0.01	2.24	64.6	94.1
	W_NMF = 0.1, W_DSD = 0.01	2.16	**66.8**	94.3
	W_NMF = 0.01, W_DSD = 0.1	2.26	64.4	93.5
	UML	**2.15**	**66.8**	**94.7**
Cas-MVSNet (Cascade)	W_NMF = 0, W_DSD = 0	2.19	66.3	93.6
	W_NMF = 0.01, W_DSD = 0.01	**2.04**	**69.6**	94.8
	W_NMF = 0.1, W_DSD = 0.01	2.12	67.1	94.8
	W_NMF = 0.01, W_DSD = 0.1	2.09	68.4	94.5
	UML	**2.04**	69.2	**94.9**

faults in both examples are well identified, while the DSD semantic maps on the right also exhibit clear object positions; moreover, as the object positions change with varying viewing angles, the contours in both semantic maps also adjust accordingly. This positional capability, from a multiview perspective, facilitates tracking of view changes and establishes correspondences between views, thereby providing guidance for depth estimation through precise view transformations.

In addition, although both the NMF and DSD branches enhance accuracy through semantic signals, they focus on different aspects. The NMF branch primarily extracts episodic semantics, while the DSD branch decomposes all features. As shown in Fig. 11, the NMF result displays block-like semantics with clear boundaries, indicating the decomposition of different components. In contrast, the semantics in the DSD result appear more diffuse and fluid. This distinction arises from the nonnegativity constraint of NMF, which allows only additive combinations and promotes independent feature extraction. On the other hand, DSD does not impose such constraints and converges to the global optimum, resulting in feature vectors that are not specific to individual objects. These observations confirm the analysis of the performance improvement reasons discussed in the previous section and reflect the rationality of our design philosophy that combines global and local information.

2) Uncertainty multitask optimization method: Likewise, in both the noncascade and cascade configurations, Table VI illustrates the results obtained by manually adjusting the weight parameters and using adaptive generated weight parameters. In each part, the first row represents the baseline model without using semantic branches, while the last row corresponds to the model incorporating uncertainty multitask loss estimation (referred to as UML) on the baseline. The intermediate rows exhibit various combinations of weights for the two semantic branches. Considering the limited discrimination capability of the unsupervised clustering method used for distinguishing similar objects in large areas such as the background, the proportions of the two branches relative to the main losses of the MVS network are kept at relatively low values. The table reveals that the optimization method using uncertainty multitask estimation achieves performance levels comparable to manually set weights and even slightly surpasses them.

TABLE VII

SENSITIVITY EXPERIMENT RESULTS FOR DIFFERENT NUMBERS OF SEMANTIC CLUSTERS IN THE PROPOSED CSC-MVS METHOD ON THE WHU TLC DATASET

Model	K_S	MAE (m)	<2.5m (%)	<7.5m (%)
MVSNet	3	2.22	65.0	93.7
	4	**2.15**	**66.8**	**94.7**
	5	2.37	62.0	93.4
	6	2.41	61.0	93.4

TABLE VIII

SENSITIVITY EXPERIMENT RESULTS FOR DIFFERENT MODELS OF FEATURE EXTRACTOR IN THE PROPOSED CSC-MVS METHOD ON THE WHU TLC DATASET

Model	Feature Extractor	MAE (m)	<2.5m (%)	<7.5m (%)
MVSNet	SwinT	2.25	64.6	93.7
	ViT	2.43	60.5	93.2
	ResNet	2.33	62.0	94.1
	VGG	**2.15**	**66.8**	**94.7**

C. Sensitivity Analysis

For the semantic-guided methods, two important parameters greatly influence the performance enhancement of MVS networks. First, the choice of the number of semantic clusters (K_S value) plays a vital role in extracting common semantic concepts across views and distinguishing different semantic categories. Second, the feature extractor model determines the quality of the extracted semantic auxiliary information. Therefore, using MVSNet as the baseline, we conducted sensitivity experiments on these two parameters separately on the WHU TLC dataset, and the results are presented in Tables VII and VIII. It is worth noting that the compared models, including SwinT [72], ViT [73], ResNet [74], and VGG [50], were pretrained on the ImageNet dataset [75]. From the two tables, it is evident that the used parameter combinations (K_S value set to 4 and VGG used as the feature extractor) are indeed the most effective at the current stage. Additionally, other choices for the K_S value and feature extractor exhibit relative robustness.

D. Generalization Testing

To validate the generalization capabilities of the CSC-MVS algorithm proposed in this study, we conducted experiments involving the transfer from the synthetic environment (i.e., LuoJia-MVS dataset) to the real environment (i.e., WHU TLC dataset) using the MVSNet framework. As seen in Table IX, for each case, the number of depth planes in test (and fine-tuning) stage is 64 (i.e., the optimal number for the WHU TLC dataset). "Baseline" signifies that conducting both training and testing exclusively in the real environment, in which the number of depth planes in training stage is also 64. "Pretraining" refers to the training from the synthetic environment, followed by fine-tuning and testing in the real environment, where the number of depth planes in pretraining stage is N_d. N_d varies progressively from the optimal number of 192 for the LuoJia-MVS dataset to the optimal number of 64 for the WHU TLC dataset. The fine-tuning process consisted of only

TABLE IX

TRANSFER RESULTS WITH VARYING DEPTH PLANE NUMBERS OF THE PROPOSED CSC-MVS METHOD FROM THE LUOJIA-MVS DATASET TO THE WHU TLC DATASET

Model	Method	MAE (m)	<2.5m (%)	<7.5m (%)
MVSNet	Baseline	**2.15**	**66.8**	**94.7**
	Pre-training with N_d=192	2.22	65.1	94.1
	Pre-training with N_d=128	2.23	65.0	94.2
	Pre-training with N_d=64	2.19	66.0	94.2

12 epochs with an initial learning rate set to 0.001, which was reduced by a factor of 2 at the 6th, 8th, and 10th epochs.

Compared to directly conducting training on real scenes, the transfer results from synthetic to real scenes are acceptable. This signifies the robust generalization abilities of CSC-MVS, enabling application transition across diverse environments. The transfer performance, furthermore, improves as the number of depth planes used during pretraining approaches the optimal number for the target scene (i.e., $N_d = 64$).

VI. CONCLUSION

In this study, we propose a new method called "CSC-MVS" that leverages clustering-based semantic information to enhance MVS networks for remote sensing 3-D reconstruction. The first contribution of our method is the semantic-guided branch comprising NMF and DSD pipelines. This branch integrates clustering techniques into the MVS network and extracts common semantics between views from global and local perspectives, thereby supplementing the additional supervision signals for depth estimation tasks. The second contribution of CSC-MVS is the uncertainty multitask optimization approach for depth estimation. By introducing the uncertainty parameter σ, the method flexibly controls the losses and adaptively combines the matching and semantic measurements. Extensive experiments on challenging remote sensing datasets, namely WHU TLC and LuoJia-MVS, demonstrate the effectiveness of CSC-MVS in the depth map reconstruction task and its general performance improvement on the current SOTA MVS networks; however, the CSC-MVS algorithm still faces certain limitations and challenges. The introduced semantic consistency method does not simplify the complexity of mainstream networks in their original structures and lacks consideration for dynamic scenes (e.g., multiview video data). Nevertheless, given the promising performance of semantic consistency in remote sensing 3-D reconstruction and its compatibility with common reconstruction outputs such as point clouds and DSMs, our future research will further explore the application of clustering methods in this field, such as designing lightweight modules that can substitute existing structures or investigating alternative extraction techniques that align better with the task.

ACKNOWLEDGMENT

The authors would also like to thank the editors and anonymous reviewers for their insightful remarks, which significantly improved this article.

REFERENCES

[1] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2006, pp. 519–528.

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

[3] E. F. Berra and M. V. Peppa, "Advances and challenges of UAV SFM MVS photogrammetry and remote sensing: Short review," in *Proc. IEEE Latin Amer. GRSS ISPRS Remote Sens. Conf. (LAGIRS)*, Mar. 2020, pp. 533–538.

[4] Y. Yao, Z. X. Luo, S. W. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. 15th European Conf. Comput. Vis. (ECCV)*, Munich, Germany, vol. 11212. Cham, Switzerland: Springer, 2018, pp. 785–801.

[5] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent MVSNet for high-resolution multi-view stereo depth inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 5520–5529.

[6] J. Liu and S. Ji, "A novel recurrent encoder–decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2020, pp. 6049–6058.

[7] K. Prokopetc and R. Dupont, "Towards dense 3D reconstruction for mixed reality in healthcare: Classical multi-view stereo vs deep learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Seoul, South Korea, Oct. 2019, pp. 2061–2069.

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

[9] J. Y. Li, X. Huang, and J. Y. Gong, "Deep neural network for remote-sensing image interpretation: Status and perspectives," *Nat. Sci. Rev.*, vol. 6, no. 6, p. 1082, Nov. 2019.

[10] T. Schöps et al., "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2538–2547.

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

[12] A. Knapitsch, J. Park, Q. Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graph.*, vol. 36, no. 4, p. 78, Jul. 2017.

[13] Y. Yao et al., "BlendedMVS: A large-scale dataset for generalized multi-view stereo networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2020, pp. 1787–1796.

[14] W. Fu, J. Ma, P. Chen, and F. Chen, "Remote sensing satellites for digital Earth," in *Manual of Digital Earth*, H. Guo, M. F. Goodchild, and A. Annoni, Eds. Singapore: Springer, 2020, pp. 55–123.

[15] J. Li, X. Huang, Y. Feng, Z. Ji, S. Zhang, and D. Wen, "A hierarchical deformable deep neural network and an aerial image benchmark dataset for surface multiview stereo reconstruction," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600812.

[16] J. Gao, J. Liu, and S. Ji, "Rational polynomial camera model warping for deep learning based satellite multi-view stereo matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, New York, NY, USA, Oct. 2021, pp. 6128–6137.

[17] X. Huang, J. Li, W. Liao, and J. Chanussot, "Information extraction from remote sensing imagery," *Geo-Spatial Inf. Sci.*, vol. 20, pp. 297–298, 2017.

[18] C. Strecha, R. Fransens, and L. Van Gool, "Combined depth and outlier estimation in multi-view stereo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 2394–2401.

[19] J.-P. Pons, R. Keriven, and O. Faugeras, "Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score," *Int. J. Comput. Vis.*, vol. 72, no. 2, pp. 179–193, Apr. 2007.

[20] P. Labatut, J. P. Pons, and R. Keriven, "Efficient multi-view reconstruction of large-scale scenes using interest points, Delaunay triangulation and graph cuts," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, Oct. 2007, pp. 504–511.

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

[22] H. Ham, J. Wesley, and H. Hendra, "Computer vision based 3D reconstruction: A review," *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 9, no. 4, p. 2394, Aug. 2019.

[23] A. Yuniarti and N. Suciati, "A review of deep learning techniques for 3D reconstruction of 2D images," in *Proc. 12th Int. Conf. Inf. Commun. Technol. Syst. (ICTS)*, Surabaya, Indonesia, Jul. 2019, pp. 327–331.

[24] J. Yan et al., "Dense hybrid recurrent multi-view stereo net with dynamic consistency checking," in *Computer Vision-(ECCV)*. Cham, Switzerland: Springer, 2020, pp. 674–689.

[25] Z. Wei, Q. Zhu, C. Min, Y. Chen, and G. Wang, "AA-RMVSNet: Adaptive aggregation recurrent multi-view stereo network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, New York, NY, USA, Oct. 2021, pp. 6167–6176.

[26] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2020, pp. 2492–2501.

[27] S. Cheng et al., "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2020, pp. 2521–2531.

[28] R. Peng, R. Wang, Z. Wang, Y. Lai, and R. Wang, "Rethinking depth estimation for multi-view stereo: A unified representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 8635–8644.

[29] X. Wang et al., "Multi-view stereo in the deep learning era: A comprehensive review," *Displays*, vol. 70, Dec. 2021, Art. no. 102102.

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

[31] Y. Cao and X. Huang, "A deep learning method for building height estimation using high-resolution multi-view imagery over urban areas: A case study of 42 Chinese cities," *Remote Sens. Environ.*, vol. 264, Oct. 2021, Art. no. 112590.

[32] S. M. Seitz and C. R. Dyer, "Photorealistic scene reconstruction by voxel coloring," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1997, pp. 1067–1073.

[33] S. S. Wong and K. L. Chan, "3D object model reconstruction from image sequence based on photometric consistency in volume space," *Pattern Anal. Appl.*, vol. 13, no. 4, pp. 437–450, Nov. 2010.

[34] H. Zhuge, "Interactive semantics," *Artif. Intell.*, vol. 174, no. 2, pp. 190–204, Feb. 2010.

[35] G. R. Yang, H. S. Zhao, J. P. Shi, Z. D. Deng, and J. Y. Jia, "Segstereo: Exploiting semantic information for disparity estimation," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, vol. 11211. Cham, Switzerland: Springer, 2018, pp. 660–676.

[36] Y. Li et al., "MFVNet: A deep adaptive fusion network with multiple field-of-views for remote sensing image semantic segmentation," *Sci. China Inf. Sci.*, vol. 66, no. 4, Apr. 2023, Art. no. 140305.

[37] J. E. Ball, D. T. Anderson, and C. S. Chan, "Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community," *J. Appl. Remote Sens.*, vol. 11, p. 54, Sep. 2017.

[38] R. T. Collins, "A space-sweep approach to true multi-image matching," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 1996, pp. 358–363.

[39] Z. Yu and S. Gao, "Fast-MVSNet: Sparse-to-dense multi-view stereo with learned propagation and Gauss–Newton refinement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1946–1955.

[40] O. D. Faugeras and F. Lustman, "Motion and structure from motion in a piecewise planar environment," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 2, no. 3, pp. 485–508, Sep. 1988.

[41] J. Yang, W. Mao, J. M. Alvarez, and M. Liu, "Cost volume pyramid based depth inference for multi-view stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4748–4760, Sep. 2022.

[42] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, pp. 234–241.

[43] K. Cho, B. V. Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proc. SSST@EMNLP*, 2014, pp. 103–111.

[44] X. J. Shi, Z. R. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. 29th Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, vol. 28, 2015, 2015.

[45] Q. Xu and W. Tao, "Multi-scale geometric consistency guided multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 5478–5487.

[46] Q. Xu, W. Kong, W. Tao, and M. Pollefeys, "Multi-scale geometric consistency guided and planar prior assisted multi-view stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4945–4963, Apr. 2023.

[47] T. Khot, S. Agrawal, S. Tulsiani, C. Mertz, and M. Hebert, "Learning unsupervised multi-view stereopsis via robust photometric consistency," 2019, *arXiv:1905.02706*.

[48] Y. Dai, Z. Zhu, Z. Rao, and B. Li, "MVS2: Deep unsupervised multi-view stereo with multi-view symmetry," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2019, pp. 1–8.

[49] B. Huang, H. Yi, C. Huang, Y. He, J. Liu, and X. Liu, "M3 VSNET: Unsupervised multi-metric multi-view stereo network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, New York, NY, USA, Sep. 2021, pp. 3163–3167.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[51] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 7482–7491.

[52] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[53] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1336–1353, Jun. 2013.

[54] E. Collins, R. Achanta, and S. Susstrunk, "Deep feature factorization for concept discovery," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, vol. 11218. Cham, Switzerland: Springer, 2018, pp. 352–368.

[55] C. Ding, X. F. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *Proc. 5th SIAM Int. Conf. Data Mining*. Philadelphia, PA, USA: SIAM 2005, pp. 606–610.

[56] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, 2000, pp. 535–541.

[57] F. R. K. Chung, *Lectures on Spectral Graph Theory*. Providence, RI, USA: American Mathematical Society, 2001.

[58] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[59] A. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. NIPS*, 2001, pp. 849–856.

[60] L. Melas-Kyriazi, C. Rupprecht, I. Laina, and A. Vedaldi, "Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 8354–8365.

[61] Y. Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in ND images," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, Vancouver, BC, Canada, Jul. 2001, pp. 105–112.

[62] W. E. Donath and A. J. Hoffman, "Lower bounds for the partitioning of graphs," *IBM J. Res. Develop.*, vol. 17, no. 5, pp. 420–425, Sep. 1973.

[63] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak Math. J.*, vol. 23, no. 2, pp. 298–305, 1973.

[64] Q. Chen, D. Li, and C.-K. Tang, "KNN matting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 869–876.

[65] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *Proc. IEEE Workshop Stereo Multi-Baseline Vis. (SMBV)*, Dec. 2001, pp. 131–140.

[66] W. van der Mark and D. M. Gavrila, "Real-time dense stereo for intelligent vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 1, pp. 38–50, Mar. 2006.

[67] T. Pei et al., "GIScience and remote sensing in natural resource and environmental research: Status quo and future perspectives," *Geography Sustainability*, vol. 2, no. 3, pp. 207–215, Sep. 2021.

[68] C. Liu, X. Huang, Z. Zhu, H. Chen, X. Tang, and J. Gong, "Automatic extraction of built-up area from ZY3 multi-view satellite imagery: Analysis of 45 global cities," *Remote Sens. Environ.*, vol. 226, pp. 51–73, Jun. 2019.

[69] X. Huang et al., "High-resolution urban land-cover mapping and landscape analysis of the 42 major cities in China using ZY-3 satellite images," *Sci. Bull.*, vol. 65, no. 12, pp. 1039–1048, Jun. 2020.

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

[71] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, "Multilabel remote sensing image retrieval based on fully convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 318–328, 2020.

[72] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[73] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[74] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[75] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

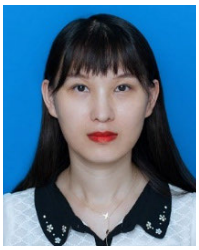
Xin Huang (Senior Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2009, working with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS).

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

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

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

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



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

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

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

Leiguang Wang received the Ph.D. degree from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, in 2009.

From 2014 to 2015, he was a Postdoctoral Researcher with the University of New Brunswick, Fredericton, NB, Canada. He is currently a Professor with Southwest Forestry University, Kunming, China, where he has been the Vice Dean of the Institute of Big Data and Artificial Intelligence since 2018. He is the author of more than 50 articles. His research interests include remote sensing image fusion, semantic segmentation, and application in forestry.