ISPRS Journal of Photogrammetry and Remote Sensing 195 (2023) 446–461

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

A general deep learning based framework for 3D reconstruction from multi-view stereo satellite images

Jian Gao, Jin Liu, Shunping Ji *

School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China

ARTICLE INFO

Keywords:
Multi-view stereo
Optical satellite images
Deep learning
Dense matching
3D reconstruction

ABSTRACT

In this paper, we propose a general deep learning based framework, named Sat-MVSF, to perform three-dimensional (3D) reconstruction of the Earth's surface from multi-view optical satellite images. The framework is a complete processing pipeline, including pre-processing, a multi-view stereo (MVS) network for satellite imagery (Sat-MVSNet), and post-processing. The pre-processing handles the geometric and radiometric configuration of the multi-view images and their cropping. The cropped multi-view patches are then fed into Sat-MVSNet, which includes deep feature extraction, rational polynomial camera (RPC) warping, pyramid cost volume construction, regularization, and regression, to obtain the height maps. The error matches are then filtered out and a digital surface model (DSM) is generated in the post-processing. Considering the complexity and diversity of real-world scenes, we also introduce a self-refinement strategy that does not require any ground-truth labels to enhance the performance and robustness of the Sat-MVSF framework. We comprehensively compare the proposed framework with popular commercial software and open-source methods, to demonstrate the potential of the proposed deep learning framework. On the WHU-TLC dataset, where the images are captured with a three-line camera (TLC), the proposed framework outperforms all the other solutions in terms of reconstruction fineness, and also outperforms most of the other methods in terms of efficiency. On the challenging MVS3D dataset, where the images are captured by the WorldView-3 satellite at different times and seasons, the proposed framework also exceeds the existing methods when using the model pretrained on aerial images and the introduced self-refinement strategy, demonstrating a high generalization ability. We also note that the lack of training samples hinders research in this field, and the availability of more high-quality open-source training data will greatly accelerate the research into deep learning based MVS satellite image reconstruction. The code will be available at https://gpcv.whu.edu.cn/data.

1. Introduction

The increasing number of high-resolution optical satellites with stereo ability has facilitated the development of three-dimensional (3D) reconstruction of the Earth's surface (de Franchis et al., 2014; Ozcanli et al., 2014) and attracted a lot of interest from researchers, in both photogrammetry and computer vision (Gao et al., 2021; Tao et al., 2004; Zhang et al., 2019). There are two different ways of capturing stereo images from space. One is to use a multiple line-scan camera mounted on a satellite, and the other is to take multiple shots of the same scene within a certain time window using a single line-scan camera. Data acquired in these two ways are widely used in practical production. For example, the Zi-Yuan-3 (ZY3) satellite belongs to the first type and the WorldView series the second.

There are some commercial software packages (Agisoft, 2022; ArcGIS, 2022; Catalyst, 2021) and free open-source solutions (de Franchis et al., 2014; Zhang et al., 2019) available to extract the 3D information from satellite stereo pairs with rational polynomial camera (RPC) parameters (Meng et al., 2007; Xiong and Zhang, 2010). Nevertheless, the existing solutions are still in the development phase, i.e., pursuing higher efficiency in large-scale reconstruction or designing more sophisticated hand-crafted matching methods to handle challenging situations. On the other hand, the advanced deep learning based multi-view stereo (MVS) methods have demonstrated their great potential in terms of accuracy and efficiency in close-range (Cheng et al., 2020; Gu et al., 2020; Yao et al., 2018) and aerial reconstruction (Liu and Ji, 2020; Yu et al., 2021) using pinhole (frame) cameras. To bridge the gap between the different imaging geometries of push-broom and pinhole cameras,

* Corresponding author.
E-mail address: jishunping@whu.edu.cn (S. Ji).

https://doi.org/10.1016/j.isprsjprs.2022.12.012
Received 17 June 2022; Received in revised form 13 December 2022; Accepted 20 December 2022
Available online 29 December 2022

we proposed a rigorous differentiable rational polynomial camera warping (RPC warping) module (Gao et al., 2021) to achieve deep MVS satellite image 3D reconstruction, without any epipolar rectification.

In this paper, we further propose a general deep learning based Framework to provide a novel solution for extracting 3D surfaces from MVS Satellite images (Sat-MVSF), which can process the triple-view images captured from a three-line camera (TLC) camera mounted on the ZY3 satellite, or any images from single line-array cameras at different times. The framework not only includes the RPC warping, which geometrically aligns the stereo images, learnable feature extraction, and inference, but also a label-free self-refinement strategy that improves the inference accuracy on unseen scenes, and pre-processing and post-processing operations that make the pipeline end-to-end.

This work is an extension of our previous conference work (Gao et al., 2021), with clear new contributions:

We propose a general and complete deep learning based framework for the MVS satellite image reconstruction task, which we believe is the first deep learning based framework developed for this task.

We provide a comprehensive description of each step that is involved when modern deep learning based technology is applied to the satellite imagery 3D reconstruction task, including pre-processing, deep feature extraction, RPC warping, pyramid cost volume construction, inference, and post-processing, with available code at https://gpcv.whu.edu.cn/data.

We provide the first comprehensive evaluation of popular software, open-source methods, and deep learning based methods on TLC and multi-view WorldView-3 images, and we show the advantages of the proposed deep learning based MVS method.

2. Related work

In this section, the two most closely related topics are reviewed, i.e., 3D reconstruction from optical stereo satellite images, and deep learning based MVS matching.

2.1. 3D reconstruction from stereo satellite images

Differing from the common close-range and aerial images, most optical satellite images are acquired through line-array scanners in a push-broom way. The RPC model, which is a general geometric model widely used in the high-resolution satellite domain, has shown negligible accuracy loss when compared to the rigorous sensor model (RSM) (Fraser and Hanley, 2003; Grodecki, 2001; Paderes, 1989). The advantages of the RPC model have thus benefited a series of studies on satellite image 3D reconstruction (Ahn et al., 2001; Fraser and Hanley, 2003; Marí et al., 2019; Toutin, 2001; Zhang et al., 2019).

Most of the methods perform epipolar rectification on the stereo satellite images, and then apply a stereo matching algorithm to obtain a high density of conjugate points. The epipolar geometry is based on the RPC model (Kuschk et al., 2014; Wang et al., 2011). The semi-global matching (SGM) (Hirschmuller, 2005) or modified SGM forms the core of many satellite image 3D reconstruction solutions. The *CATENA* (Krauβ et al., 2013) uses SGM with distributed optimization to generate a high-resolution digital surface model (DSM) automatically. RPC Stereo Processor (RSP) (Qin, 2016) includes rectification, geo-referencing, point cloud generation, pan-sharpening, DSM resampling, and *ortho*-rectification, to generate DSMs and orthophotos. (Rupnik et al., 2018) applied MicMac (Rupnik et al., 2017) to dense matching of very-high-resolution (VHR) satellite images and proposed a multi-view depth map fusion method through a multi-directional dynamic programming approach to complete DSM generation. Most of the commercial software packages follow similar technical routes to those described above. For example, CATALYST (Catalyst, 2021), which is a brand developed by PCI Geomatics, uses a multi-view SGM technique to extract 3D information from satellite and aerial images. The Ortho Mapping tool in ArcGIS (ArcGIS, 2022) provides a series of tools to derive high-resolution DSMs from aerial or satellite stereo pairs. Agisoft Metashape (Agisoft, 2022) is also capable of processing panchromatic and multispectral satellite images, since Version 1.7 (2021).

There are also a few methods that rectify stereo images but approximate the push-broom geometry of small cropped image tiles by a pinhole model, and then perform standard stereo matching (e.g., more global matching (MGM) (Facciolo et al., 2015)). Satellite Stereo Pipeline (S2P) (Franchis et al., 2014) is a typical example of this.

The other type of methods use plane sweeping to avoid epipolar resampling, to reconstruct the 3D structure from multi-view satellite images. However, the method proposed in (Zhang et al., 2019) has to approximate the RPC model as a perspective camera model, and then adopts the plane sweeping based MVS of COLMAP (Schönberger and Frahm, 2016; Schönberger et al., 2016) for 3D reconstruction from satellite images.

To the best of our knowledge, the current solutions for 3D reconstruction from satellite images are all based on conventional methods.

2.2. Deep learning based multi-view stereo methods

The plane-sweeping algorithm (Collins, 1996) is designed for depth-wisely aligning corresponding points in the conventional MVS methods for pin-hole cameras, which is also the strategy used in modern deep learning based MVS methods. It is quite similar to the well-known vertical line locus (VLL) method (Fan et al., 2007; Helava, 1988; Zhang and Gruen, 2006) in photogrammetry. MVSNet (Yao et al., 2018) is one of the earliest end-to-end learnable solutions that infers the depth map from multi-view images, where deep features extracted from the images are used to replace pixel intensity or shallow hand-crafted features and the plane sweeping cost volume is constructed through the differentiable homography warping. However, the memory requirement of 3D convolutional layers grows cubically with the resolution of the input volume, which limits the sampling number along the depth search range and the resolution and size of the input images. Cas-MVSNet (Gu et al., 2020), UCS-Net (Cheng et al., 2020) and CVP-MVSNet (Yang et al., 2020) introduce a pyramid matching structure based on MVSNet. At the top of the pyramid, a coarse depth map is estimated with coarse depth intervals. Then, in the following stage, the depth search range is narrowed to a local buffer around the coarse estimated depths, and a thin cost volume is constructed with a few hypothetical depth planes but a higher depth-wise sampling rate. Some other methods adopt a recurrent convolutional structure, which depth-wise processes the cost volume instead of processing it as a whole, to reduce the memory requirement. For example, R-MVSNet (Yao et al., 2019) replaces the 3D CNN with the stacked convolutional variant of the gated recurrent unit (GRU), i.e., ConvGRU. This recurrent convolutional structure treats the depth search direction as the temporal dimension, regularizes the 2D cost maps sequentially across the depths, and uses hidden states to record the contextual information. Subsequently, the RED-Net (Liu and Ji, 2020) and DH-RMVSNet (Yan et al., 2020) methods were introduced, which combine the ConvGRU and ConvLSTM with a 2D U-Net structure, respectively, to form a multi-scale recurrent regularization structure, to aggregate the geometric and contextual information in the depth direction. The recurrent regularization structure is a trade-off between memory and efficiency, and the memory requirement is independent of the number of depth planes. A finer division of the search range is therefore possible, but the run time increases linearly, due to the sequential processing of the cost maps. Some more recent approaches (Ding et al., 2022; A. Yu et al., 2021; Zhang et al., 2021) focus on enhancing the feature extraction module by introducing the attention mechanism. Some other approaches (Wei et al., 2021; Xu et al., 2022; Zhang et al., 2022) stress that the pixel-wise visibility is crucial for the MVS problem because the occlusions and noises of different views create inconsistent information. Specifically, Vis-MVSNet (Zhang et al., 2022) and PVS-Net (Xu et al., 2022) use a 3D CNN module to learn the visibility map from the two-view cost volume while AA-RMVSNet (Wei et al.,

2021) introduces the intra-view and inter-view attention modules for adaptive cost aggregation. There are a few methods that have abandoned the structured cost volume, such as Patchmatch-Net (Wang et al., 2021), which introduces the idea of the traditional Patchmatch (Bleyer et al., 2011) into an end-to-end MVS network, where the memory consumption and run time are significantly reduced by learnable adaptive propagation and cost aggregation.

Recent learning based MVS methods (Ding et al., 2022; Wei et al., 2021; Xu et al., 2022; Zhang et al., 2022) have shown their impressive performance on well-known benchmarks (Jensen et al., 2014; Knapitsch et al., 2017; Schops et al., 2017) in computer vision community. However, few deep learning based MVS methods have been applied to 3D reconstruction from optical satellite images. This is due to the fact that the RPC model differs significantly from the perspective geometry model, which most of the deep learning-based methods are based on. In addition, it is more challenging to acquire large-scale and high-quality ground truth for training, compared to close-range images.

3. Deep learning based framework for MVS satellite imagery

We propose a novel, complete, and practical deep learning based MVS reconstruction Framework for Satellite images, named Sat-MVSF. The whole framework consists of a non-learning pre-processing module (Section 3.1), a non-learning post-processing module (Section 3.7), and a learning-based MVS inference module which we call Sat-MVSNet. Sat-MVSNet is divided into six parts: deep feature extraction (Section 3.2), RPC warping (Section 3.3), pyramid cost volume construction (Section 3.4), inference (Section 3.5), supervised training loss (Section 3.6), and optional self-refinement (Section 3.8). The inputs of the Sat-MVSF framework are multi-view satellite images as well as rational polynomial coefficients, and the output is a DSM. The framework can process multi-view images in the inference stage, whereas the deep learning model can be unrestrictedly trained on stereo, TLC, multi-view images, or their combination, and epipolar resampling is not required. A multi-scale structure is adopted to accommodate the huge Earth surface elevation differences, and to boost efficiency. The overall Sat-MVSF framework is illustrated in Fig. 1.

3.1. Pre-processing

3.1.1. Selection of stereo pairs

The accuracy of MVS reconstruction is related to factors such as the intersection angle, incidence angle, resolution difference, and scene changes. The TLC on the ZY3 satellite acquires multi-view images with well-designed intersection angles and a similar resolution at almost the same time. However, the WorldView-3 images, such as those in the MVS3D dataset (Bosch et al., 2016), are acquired from different locations at different times and seasons. The selection of proper stereo pairs is required for better reconstruction. Following the idea in (Facciolo et al., 2017), image pairs with intersection angles between α and β and maximum incidence angles less than γ are selected and then sorted in order of time difference. The first-ranked image pairs are processed first.

3.1.2. Bundle adjustment

Bundle adjustment with affine transformation in the image space (Fraser and Hanley, 2003) can be used when the acquired RPC parameters are biased. Scale-invariant feature transform (SIFT) (Lowe, 2004) and similar features can be extracted from each image and matched between stereo pairs, to obtain the conjugate points. If there are no absolute control constraints (e.g., control points), one image is selected as the reference, and affine transformation is applied to the other image. If there are absolute control constraints, affine transformation is applied to all the images. After error compensation, for each image, based on the affine transformation and the original RPC model, a virtual control point grid is first constructed and then fitted by a new RPC model. For more details, we refer the reader to (Tao and Hu, 2001).

3.1.3. Radiometric processing

Satellite images are characterized by high dynamic range (HDR), and the intensity distribution of pixels typically shows a long-tailed distribution (Zhang et al., 2019), which is not conducive to feature matching and dense matching. We use a τ% linear stretch to enhance the contrast and suppress the areas that are too dark or too bright to obtain low dynamic range (LDR) images with well-distributed intensity. Firstly, histogram statistics are calculated for the whole image. The pixels whose values are located in the lowest τ% in the histogram are then set to 0, the pixels in the highest τ% are set to 255, and the remaining pixels are

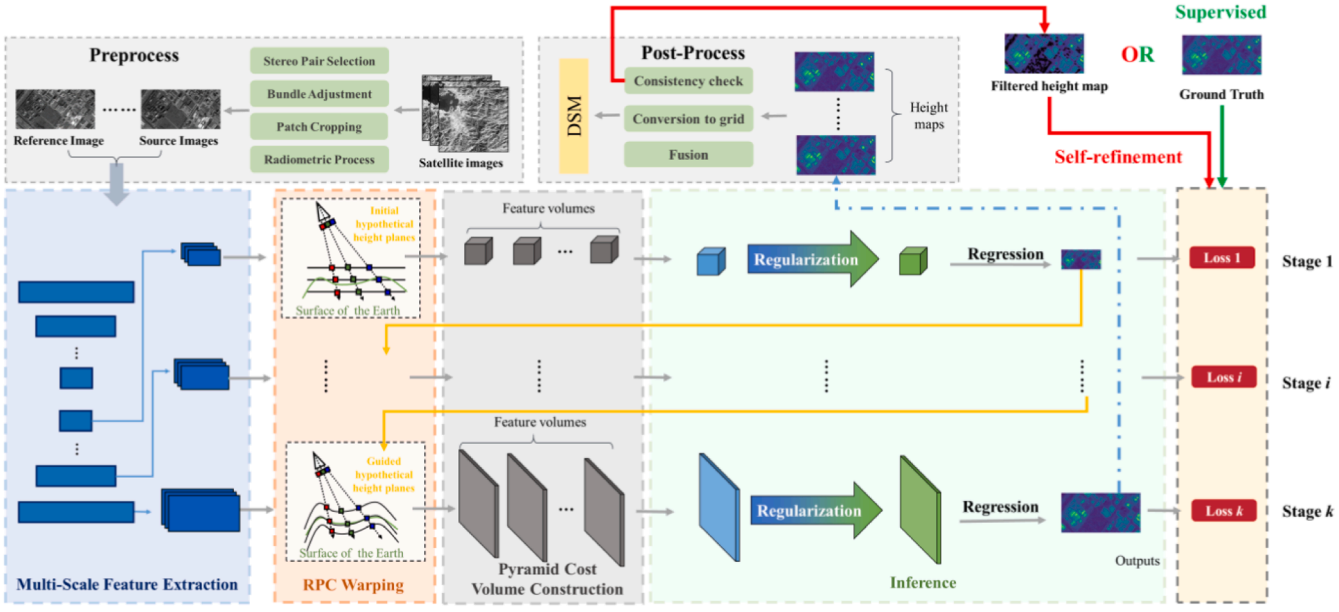

Fig. 1. Sat-MVSF for 3D Earth surface reconstruction from optical satellite images using a deep learning based multi-view stereo method, which consists of eight modules: 1) pre-processing; 2) multi-scale deep feature extraction; 3) RPC warping; 4) pyramid cost volume construction; 5) inference; 6) supervised training loss; 7) post-processing; and 8) optional self-refinement.

linearly stretched to 0 to 255. This step is optional, depending on the actual quality of the images.

3.1.4. Patch cropping

The huge size of satellite images makes it impossible for them to be processed as a whole, due to the limited GPU capacity. It is therefore necessary to crop the images into patches with a reasonable overlap for dense matching and subsequent restoration. In addition, cropping multi-view patches as the same size helps the batch processing in a deep learning framework.

In this work, we define the boundary of the DSM product in advance, and then divided the DSM into uniform blocks. Each block is then projected onto the multi-view images through the (refined) RPC model, and a minimum outside rectangle is calculated for each view. The largest minimum outside rectangle is then used to crop images from different views into patches. This cropping scheme ensures that the multi-view image patches have the same size and sufficient overlap.

3.2. Deep feature extraction

We use a multi-scale structure as the extractor to extract multi-scale deep features from different spatial resolutions. The feature extractor adopts an *n*-way Siamese structure, which can take arbitrary n views as input for the MVS matching (Fig. 2a). It extracts the features of each image by a ladder-structured CNN (Fig. 2b), with the weights shared among the different views. Here, for each input image, the module outputs three-stage feature maps $\{F_1, F_2, F_3\}$, with the size of {1/16,1/4,1} of the input image size and channel numbers of {64, 32, 8}, respectively, to form a feature pyramid.

3.3. RPC warping

A homography matrix can describe the relationship between a stereo image pair with known camera parameters and a set of hypothetical parallel depths in the frustum space in modern deep MVS methods (Yao et al., 2018). However, for a push-broom camera, the relationship goes far beyond matrix operations. To align multi-view satellite images (or image feature maps) in 3D space, we propose a differentiable warping module based on the RPC model, which ensures that the advanced deep learning based MVS methods can be directly introduced into satellite image 3D reconstruction, without epipolar rectification or pinhole camera model fitting.

3.3.1. RPC model

The RPC model, including a forward form (Eq. (1)) and an inverse form (Eq. (2)), uses third-order rational polynomials to express the transformation between the image and object coordinates of a point.

$$\begin{cases} samp_n = \dfrac{P_1^{fwd}(lat_n, lon_n, hei_n)}{P_2^{fwd}(lat_n, lon_n, hei_n)} \\ line_n = \dfrac{P_3^{fwd}(lat_n, lon_n, hei_n)}{P_4^{fwd}(lat_n, lon_n, hei_n)} \end{cases} \tag{1}$$

$$\begin{cases} lat_n = \dfrac{P_1^{inv}(samp_n, line_n, hei_n)}{P_2^{inv}(samp_n, line_n, hei_n)} \\ lon_n = \dfrac{P_3^{inv}(samp_n, line_n, hei_n)}{P_4^{inv}(samp_n, line_n, hei_n)} \end{cases} \tag{2}$$

where (lat_n, lon_n, hei_n) is the normalized world coordinates of an object point, and ($samp_n$, $line_n$) is the normalized image coordinates of the corresponding image point. P represents a cubic polynomial. The different subscripts (1, 2, 3, and 4) represent different polynomials in the formulas.

Many satellite image products provide only the RPC parameters in the forward form. In this case, the RPC parameters in the inverse form can be obtained by fitting. The fitting process is divided into two steps. Firstly, a virtual control point grid is constructed through the forward RPC model. Secondly, the virtual control point grid and the corresponding image points are fitted by the least-squares method using the inverse model.

3.3.2. Differentiable RPC warping

RPC warping implements a differentiable warping of an image to be matched (here we call it the source image) to the reference image through a set of hypothetical parallel height planes in the object space, as shown in Fig. 3. We divide the whole process into three subparts: 1) projecting the source image to 3D space with Eq. (2); 2) reprojecting to the reference with Eq. (1); and 3) resampling. The critical point is that we need to discover a rapid transformation for the whole image to replace the commonly used pixel-wise calculation with the RPC model, as the extremely low efficiency of the latter is unsuitable for a modern deep learning based MVS method. Since the basic forms of the forward and inverse RPC models are the same, without loss of generality, we take the projection from 3D object space to the reference image as an example to demonstrate the RPC warping in the following.

Firstly, the quaternary cubic form (QCF) and element-wise division are used to implement Eq. (1). The numerator and denominator have the same mathematical form, and can both be expressed as: $f(x_1, x_2, x_3, x_4) = \sum a_{ijk}\, x_i x_j x_k$ (i,j,k ∈{1,2,3,4}), where (x_1, x_2, x_3) are the normalized coordinates, $x_4 \equiv 1$, and a_{ijk} are the polynomial coefficients. The vector $\mathbf{X} = (x_1, x_2, x_3, x_4)$ is actually the homogeneous coordinates of the normalized coordinates. A coefficient tensor $\mathbf{T} \in \mathbb{R}^{4\times4\times4}$ is then used to record the coefficients. The relationship between the element $\mathbf{T}^{(i,j,k)}$ at position

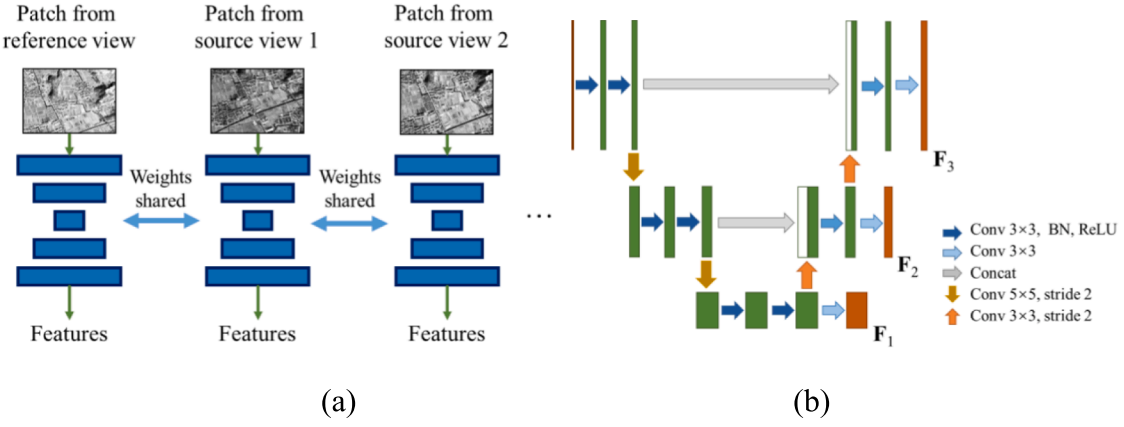

Fig. 2. The structure of the deep feature extraction module. (a) The *n*-way Siamese feature extractor. (b) The network structure and parameters of the extractor.

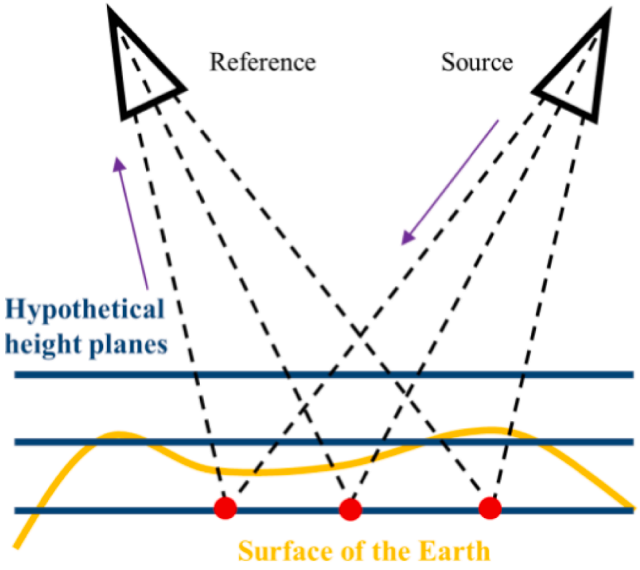

Fig. 3. RPC warping achieves differentiable warping of images or deep features from the source image to the reference image through a set of hypothetical height planes in the 3D space.

(i, j, k) in the coefficient tensor and a_{ijk} in the polynomial is: when i, j, and k are equal, $\mathbf{T}^{(i,j,k)} = a_{ijk}$; when only two of them are equal, $\mathbf{T}^{(i,j,k)} = a_{ijk}/3$; when none of them are equal, $\mathbf{T}^{(i,j,k)} = a_{ijk}/6$, as shown in Fig. 4. Therefore, f can be expressed as shown in Eq. (3) to implement the tensor expression for the numerator or denominator of Eq. (1), where $\mathbf{X}^{(i)}$ represents the i-th component of vector **X**.

$$f(X) = \sum_{i,j,k=1}^{4} T^{(i,j,k)} X^{(i)} X^{(j)} X^{(k)} \tag{3}$$

After the numerator and denominator are computed, element-wise division is performed. For resampling, we use differentiable bilinear interpolation. Finally, the source image can be differentially warped into the reference through a set of hypothetical parallel height planes for the object space, as shown in Fig. 3.

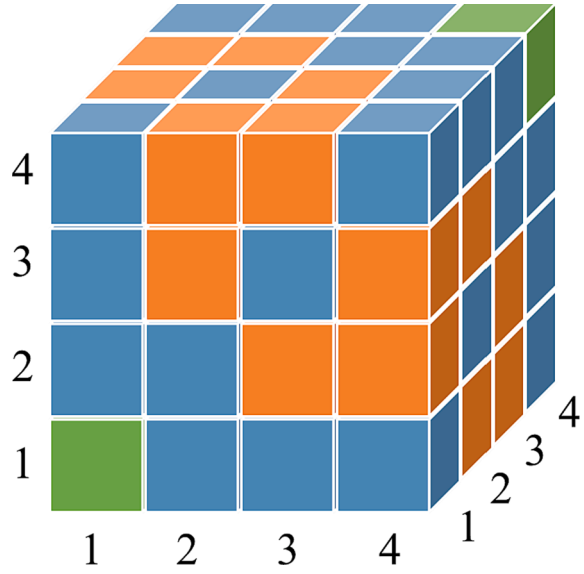

Fig. 4. The coefficient tensor $\mathbf{T} \in \mathbb{R}^{4\times4\times4}$ records the polynomial coefficients in the numerator and denominator of the RPC model. In the green cells, $\mathbf{T}^{(i,j,k)} = a_{ijk}$, $\mathbf{T}^{(i,j,k)} = a_{ijk}/3$ in the blue cells, and $\mathbf{T}^{(i,j,k)} = a_{ijk}/6$ in the orange cells. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Furthermore, to achieve batch efficient RPC warping for the whole image or the feature maps, we extend the QCF. The extended quaternary cubic form (EQCF) is shown in Eq. (4), where u and v represent the row and column number in the image or feature map, c is the channel number, b represents the batch size, and d is the index of the height plane. The tensor-based EQCF performs the coordinate transformation of all the pixels of a source image to the reference through all the hypothetical height planes simultaneously.

$$f^{(b,d,c,u,v)}(\boldsymbol{X}) = \sum_{i,j,k=1}^{4} \boldsymbol{T}^{(b,i,j,k)} \boldsymbol{X}^{(b,d,c,u,v,i)} \boldsymbol{X}^{(b,d,c,u,v,j)} \boldsymbol{X}^{(b,d,c,u,v,k)} \tag{4}$$

3.4. Pyramid cost volume construction

3.4.1. Cost metrics

Deep feature maps are warped from different source views to the reference view using RPC warping through a series of height planes to build the feature volumes. To aggregate the feature volumes from different views $\{F_v\}V$ v = 1 into one that is easy for representation and learning, we follow (Yao et al., 2018) and calculate the variance of the element values at the same position in the feature volumes to form a single cost volume C.

3.4.2. Pyramid cost volumes

Considering the huge elevation differences existing on the Earth's surface, we imbed a k-stage pyramid structure (Gu et al., 2020) in Sat-MVSNet for coarse-to-fine matching, which exactly corresponds to the 2 × down-sampled feature pyramid described in Section 3.2, i.e., k = 3. In stage i, a series of hypothetical height planes are constructed at a fixed interval I_i within the elevation search range SR_i (with the low and high boundaries SR_{Li} and SR_{Hi}), based on which a cost volume C_i is then constructed using the cost metrics mentioned in Section 3.4.1.

In the first stage (i = 1), the elevation search range SR_1 covers the maximum and minimum heights of the ground, which are equidistantly sampled as the initial hypothetical height planes, given the number of planes N_1. The maximum and minimum elevations can be calculated from the normalized parameters in the RPC model. Alternatively, several open-source global digital elevation model (DEM) products (e.g., the Shuttle Radar Topography Mission (SRTM) DEM (Jarvis et al., 2008) or the ASTER GDEM (Tachikawa et al., 2011)) could be used as information sources.

Based on the prediction of the previous stage$\widehat{h}_{i,x}$, the following stages i + 1 (i > 0) only need to refine the height prediction locally. For pixel ×, a search space centered on the predicted elevation value of the previous stage is used for stage i + 1.

$$SR_{i+1,x} = \left[\widehat{h}_{i,x} - -\frac{1}{2}\cdot I_{i+1}\cdot N_{i+1}, \widehat{h}_{i,x} + \frac{1}{2}\cdot I_{i+1}\cdot N_{i+1}\right] \tag{5}$$

where N_{i+1} and I_{i+1} are, respectively, the preset number of hypothetical planes and the sampling interval in stage i + 1.

Empirically, we gradually reduce the values of N_i and I_i as i increases, and hence the search range is more tightly restricted, to reduce the consumption of computational resources.

3.5. Inference

The inference is pyramidally executed on the pyramid cost volume. At each stage, the height map is inferred from the cost volume through regularization and regression. In the training phase, the multi-stage inferred height maps are involved in supervision, as shown in Section 3.6; while in the prediction phase, the inference result of the finest stage k is output as the final height map.

3.5.1. Regularization

The role of the so-called regularization module in some deep learning based MVS methods is similar to that of the cost aggregation (Scharstein and Szeliski, 2002) in the traditional MVS methods, in that it sequentially aggregates the information of the neighborhood features from the cost volume through a set of neuron layers. Most of the recent deep learning based MVS methods (Cheng et al., 2020; Gu et al., 2020; Yao et al., 2018) adopt layers of 3D convolutions for the regularization, but the memory requirement grows cubically with the resolution of the input volume, which hinders their use on large-size satellite images.

In the proposed Sat-MVSF, the recurrent encoder-decoder (RED) structure (Liu and Ji, 2020), which combines ConvGRUs and 2D U-Net, is adopted to regularize the cost maps. The RED structure learns the features in the spatial direction via 2D CNNs and aggregates the geometric and contextual information in the search direction (i.e., the depth direction) via ConvGRUs, thus avoiding the 3D convolution calculation. Specifically, along the height direction in the object space, the cost volume of each stage is split into a series of cost maps, which are aggregated through the RED structure and finally restacked into a cost volume. Hence, a large-size image with a huge depth range can be processed with limited GPU memory, which is especially suitable for Earth surface reconstruction from satellite images. As shown in Fig. 5, the four-scale feature pyramid maps produced by the encoder of U-Net are respectively regularized by the four-scale ConvGRUs. The regularized features are then summed with those of the same scale produced by the decoder, and are used as the input of the next deconvolutional layer in the decoder. Each scale of ConvGRU has a state, which records the information of the current input cost map, and serves as the initial state value of the next input cost map. For more information about the RED structure, we refer the reader to (Liu and Ji, 2020). Nevertheless, other efficient structures can also be used to replace the RED structure for regularization.

3.5.2. Regression

The height map is inferred from the aggregated cost volume. We follow the regression method proposed in (Yao et al., 2018), which is commonly used in most of the deep learning based MVS methods (Cheng et al., 2020; Gu et al., 2020). A softmax operation is applied along the elevation direction to convert the aggregated cost volume to a probability volume **P**. The value of **P** represents the probability of this pixel being on the current height plane j. The final estimation is then obtained by taking the sum of each height value h weighted by the normalized probability:

$$\widehat{h}_x = \sum_{j=1}^{N} h_{j,x} \cdot \boldsymbol{P}_{j,x} \tag{6}$$

where $h_{j,x}$ and $\mathbf{P}_{j,x}$ denote the height value and the probability of pixel × at the j-th height hypothetical plane, respectively. This process is fully differentiable and is able to regress a continuous estimate.

3.6. Supervised training loss

If the ground truth is available, we can train the network in a supervised manner by minimizing the smooth L1 loss between the estimated height values and the ground truth. The loss L_i at stage i is given by:

$$L_i = \begin{cases} \sum_{x \in Valid} 0.5\left(h_{i,x} - \widetilde{h}_{i,x}\right)^2 & if \left|h_{i,x} - \widetilde{h}_{i,x}\right| < 1 \\ \sum_{x \in Valid} \left(\left|h_{i,x} - \widetilde{h}_{i,x}\right| - 0.5\right) & otherwise \end{cases} \tag{7}$$

where Valid refers to the set of valid grid cells in the ground truth, and $h_{i,x}$ and $\widetilde{h}_{i,x}$ respectively represent the value of the predicted height map and the ground truth at point × in stage i. The total loss is defined as a weighted sum of the multi-stage L_i:

$$Loss = \sum_{i=1}^{N} w_i \cdot L_i \tag{8}$$

where L_i denotes the loss of stage i, and w_i denotes the weight of the same stage. In practice, the weights are set to $w_i = \{0.5, 1, 2\}$, where 2 corresponds to the finest scale.

3.7. Post-processing

3.7.1. Geometric consistency check

In practice, a set of *n*-view patches is fed into Sat-MVSNet n times. At each time, a patch is chosen as the reference (the rest as the source) in turn to obtain n height maps. The fact that the height values of an object point in the n height maps should be consistent provides an idea for the detection and removal of the incorrect matches.

A geometric consistency checking scheme similar to that in (Yao et al., 2018) is adopted here. Based on the height predicted for the reference view, a point p_1 in the reference image is transformed to the point p_2 in the source image through the inverse form of the reference RPC model and the forward form of the source RPC model. Then, based

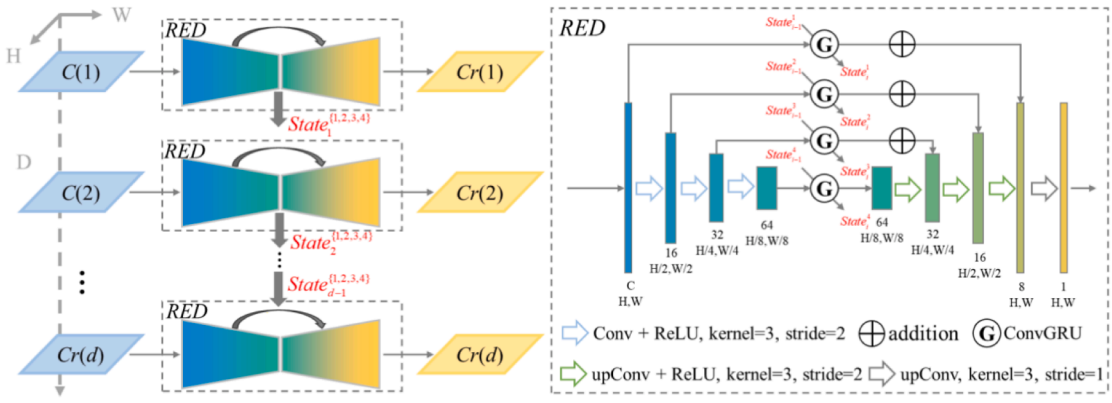

Fig. 5. Regularization with the Recurrent Encoder and Decoder (RED) structure. C(i) represents the i-th cost map after the cost volume is split along the elevation direction, and State{1,2,3,4} i is the set of the hidden states of the convolutional GRU. RED takes the C(i) and State{1,2,3,4} i-1 as inputs and outputs the regularized cost map Cr(i) and the updated State{1,2,3,4} i. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

on the height predicted from the case of the source view playing the role of "reference", point p_2 is reprojected to point p_3 on the reference image. If $||p_3 - p_1||_2 < \psi$, the two views are geometrically consistent. For the predicted result for a pixel on the reference view, the match is valid if there are at least z source views geometrically consistent with the reference; otherwise, it is removed as an outlier.

3.7.2. Conversion from point cloud to DSM

The reference height map is projected into the object space to produce an unstructured point cloud through the RPC model. The point cloud is then converted into a DSM in the following way. Firstly, the coverage of the point cloud in the object space is divided into a uniform grid in the x-direction and y-direction at a certain interval. The points in the point cloud are then projected orthogonally onto the grid cells. Finally, the maximum elevation value of all the points in each grid cell is kept, and the grid cells without any point projected in are set as invalid.

3.7.3. Fusion

DSM fusion is used to remove errors and increase the completeness when there are multiple overlapped DSMs produced from different data sources. By assuming that the error between the multiple produced DSMs and the ground truth is normally distributed, outlier detection and rejection are performed using the median absolute deviation (MAD) (Leys et al., 2013), after which the remaining valid DSM predictions are averaged as the final estimation.

3.8. Self-refinement

Differences such as the sensor type, imaging conditions, times, and geographical locations lead to large differences in both the radiometric and geometric configurations of stereo satellite images. In addition, collecting a large-volume satellite dense matching dataset which contains a wide variety of scenes is not a simple task. Both problems challenge the generalization ability of supervised deep learning based MVS methods. It is in fact likely that the datasets already collected are not sufficient to train a model that performs perfectly in all real-world scenarios.

This concern drives us to introduce a strategy for adapting the pretrained model to the target real-world data using only the information embedded in the unlabeled target data. The strategy is a simple but effective self-refinement scheme. This scheme can be divided into the following steps as shown in Fig. 6. Firstly, train the network on a publicly available dataset to obtain a pretrained model. Secondly, load the pretrained model to predict the height map of the target satellite data. Thirdly, check the predicted height map with the relative geometric consistency scheme described in Section 3.7.1, which removes errors and mines those correctly matched points as pseudo-labels. Finally, the target images and the corresponding pseudo-labels are refed into the pretrained model for adaptive refinement. The self-refinement ensures that the deep learning model can be smoothly applied to various real-world scenarios, thus greatly improving the practicality of the proposed Sat-MVSF framework.

Because the RPC warping and the homography warping are both differentiable geometric transformations without any learnable parameters, a pretrained Sat-MVSNet on close-range (e.g., DTU(Jensen et al., 2014)) or aerial (e.g., WHU-MVS(Liu and Ji, 2020)) images can be applied directly to satellite images. Nevertheless, the self-refinement in MVS matching is somehow different from that in a semantic segmentation task (Zheng and Yang, 2021). In the former, the pseudo labels we obtained are strictly checked by the multi-view geometry consistency. Therefore, the refinement training process is essentially a fine-tuning with semi-dense ground truth. But in the latter, the pseudo labels are only a probabilistic guess given by the pretrained model, which are noisy.

4. Experiments

There is currently a serious lack of training samples for deep learning based satellite MVS matching. To the best of our knowledge, the only relevant available datasets are the WHU-TLC (Gao et al., 2021), MVS3D

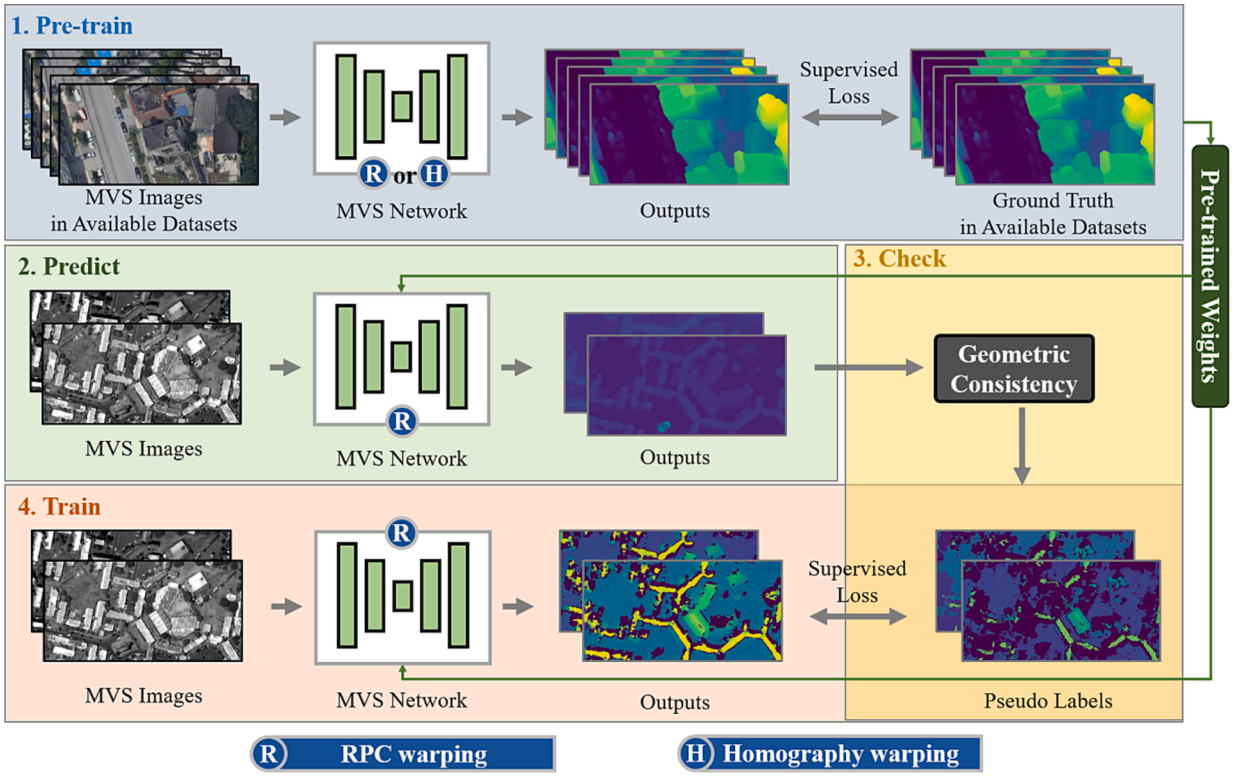

Fig. 6. Label-free self-refinement strategy.

(Bosch et al., 2016), and US3D (Bosch et al., 2019) datasets.

The WHU-TLC dataset provides sufficient samples for training, and is used as the primary dataset in the experiments described in Sections 4.2 and 4.3. The size of the MVS3D dataset is too small for it to be used for training. Thus, we only use it for the assessment of the generalization ability, as described in Section 4.4. The US3D dataset is aimed at the joint task of semantic segmentation and 3D reconstruction, where the scene variations between the stereo image pairs are not suitable for high-accuracy MVS reconstruction, so we do not use this dataset. Finally, we give some visual examples of DSM production using the proposed Sat-MVSF framework in Section 4.5.

4.1. Evaluation metrics

In this paper, we adopt the following metrics to evaluate the quality of DSMs:

1) Root-mean-square error (RMSE): the standard deviation of the residuals between the ground truth and the estimation:

$$RMSE = \sqrt{\frac{\sum_{(i,j)\in G\cap\widetilde{G}}\left(h_{ij}-\widetilde{h}_{ij}\right)^2}{\sum_{(i,j)\in G\cap\widetilde{G}} I\left((i,j)\in G\cap\widetilde{G}\right)}} \tag{9}$$

where G and $\widetilde{G}$ represent the valid grid cells in the estimated DSM and ground truth, h_{ij} and $\widetilde{h}_{ij}$ refer to the height value of the estimation and ground truth in the grid cell in row i and column j, and I(A) represents the Iverson bracket, which means 1 if A is true and 0 otherwise.

2) Mean absolute error (MAE): the average of the L_1 distance over all the grid units between the ground truth and the estimated DSM:

$$MAE = \frac{\sum_{(i,j)\in G\cap\widetilde{G}}|h_{ij}-\widetilde{h}_{ij}|}{\sum_{(i,j)\in G\cap\widetilde{G}} I\left((i,j)\in G\cap\widetilde{G}\right)} \tag{10}$$

3) Percentage of accurate grids in total (PAG): PAG is a completeness metric that takes into account accuracy as well. It defines what is an accurate grid cell and then calculates the completeness using the number of accurate grid cells. The accurate grid cells refer to those where the L_1 distance error between the ground truth and estimated DSM is below a certain threshold α.

$$PAG_\alpha = \frac{\sum_{(i,j)\in G\cap\widetilde{G}} I\left(\left|h_{ij}-\widetilde{h}_{ij}\right|<\alpha\right)}{\sum_{(i,j)\in G\cap\widetilde{G}} I\left((i,j)\in\widetilde{G}\right)} \tag{11}$$

4) Median height error (Median): the median value of the absolute error between the estimation and ground truth in the valid grid cells.

$$Median = \underset{(i,j)\in G\cap\widetilde{G}}{median}(|h_{ij}-\widetilde{h}_{ij}|) \tag{12}$$

5) Time: The run time required for the entire process from importing a set of multi-view images to the completion of DSM production.

4.2. Performance on the WHU-TLC dataset

The WHU-TLC dataset (Gao et al., 2021) is a large-scale MVS satellite image dataset built to advance the development of satellite dense matching and 3D reconstruction. There are 173 scenes in the dataset, including 127 for training and 46 for testing. The triple-view images were acquired from the TLC mounted on the ZY3-02 satellite, and have been cropped to patches of 5120 × 5120 pixels. The ground resolutions of nadir and 22° forward and backward view are 2.1 m and 2.7 m, respectively. The RPC models have been aligned with the provided DSMs, and the DSMs are stored as regular grids with a resolution of 5 m under the WGS-84 geodetic and the Universal Transverse Mercator (UTM) projection coordinate systems.

The training set of the WHU-TLC-Version 2 dataset is used to train Sat-MVSNet on mainstream GPUs. A training sample includes a multi-view image patch with the size of 768 × 384 pixels, the RPC parameters of the patch, and the height map of the reference view. However, for the testing, we use the full-size image of 5120 × 5120 pixels to evaluate the performance of the proposed Sat-MVSF framework as well as other methods.

4.2.1. Implementation details

A. Training Details.

Sat-MVSNet is implemented in PyTorch (Paszke et al., 2019) and trained on the training set of the WHU-TLC-V2 dataset with a single NVIDIA TITAN RTX GPU (24 GB). The hyper-parameter settings are as follows. In the training phase, the batch size is set to 1, and RMSProp is selected as the optimizer. The network is trained for 35 epochs with an initial learning rate of 0.001, which is downscaled by a factor of 2 after the 10th epoch. Three-stage hierarchical matching is adopted to infer the coarse-to-fine height maps. For the TLC images, the view number of the input images n is fixed to 3. The numbers of hypothetical height planes and the corresponding interval are set to {64, 32, 8} and {(SR_{H1}–SR_{L1})/64, 5 m, 2.5 m}, respectively.

B. Comparison Methods.

S2P (de Franchis et al., 2014). Satellite Stereo Pipeline (S2P) is an automatic and modular stereo pipeline for push-broom images. The images are divided into small tiles and processed in parallel with multiple processes, to improve the efficiency. In our experiment, the tile size is set to 640 pixels, a fixed elevation range is used for the disparity determination, the threshold of outlier removal in the fusion step is chosen as 25 m (about ten times the image resolution as suggested by the original S2P), and the dense matching method in S2P is the default MGM algorithm (Facciolo et al., 2015). The built-in height map outlier cleaning is not allowed here (otherwise the completeness rate is extremely low). The other settings remain the default settings.

SDRDIS (Sdrdis, 2016)**.** SDRDIS won 1st place in the Explorer Contest and 3rd place in the Master Contest of the IARPA Multi-View Stereo 3D Mapping Challenge. It relies on the tools provided in the NASA Ames Stereo Pipeline (Moratto et al., 2010) for satellite image processing and OpenCV (Bradski, 2000) for dense matching. In the experiment, bundle adjustment is disabled because the RPC parameters in the dataset are already accurate. The semi-global block matching (SGBM) algorithm is chosen, with the maximum disparity set to 288, and all the other parameters set to the default values. Since this solution does not contain an algorithm for generating a DSM, the method described in Section 3.7.2 is used after the point clouds were generated.

Adapted COLMAP (Zhang et al., 2019). Adapted COLMAP utilizes the reconstruction pipeline from the computer vision community for 3D reconstruction, where the RPC model of the local satellite image patches is fitted by a pinhole camera model. In the experiment, bundle adjustment is disabled and all the other parameters were set to the default values. Adapted COLMAP itself is not capable of handling large images. Therefore, the images are cropped (see Section 3.1.4) and then fed into Adapted COLMAP to obtain a series of small DSM blocks that are eventually stitched together to form a final DSM.

ArcGIS (ArcGIS, 2022). The Ortho Mapping module of ArcGIS 10.8 is used in the experiment. The following processes are performed sequentially: create mosaic dataset, add rasters to mosaic dataset, build stereo model, generate point cloud, and interpolate from point cloud. The matching method is set to SGM, while triangulation is used for the interpolation and a filter with a Gaussian kernel of size 5 × 5 for the smoothing. The other parameters are set to the default values.

CATALYST (Catalyst, 2021). The available trial version of CATALYST (previously known as PCI Geomatica) Professional Version 2222, SP4, 2021–09-21 is used in the experiment. The SGM algorithm is selected, the scale for the epipolar resampling is set to 1, and all the other parameters remain at the default values.

Metashape (Agisoft, 2022). The software we acquired is the trial version of Agisoft Metashape Professional 1.75. For high-quality reconstruction, we select the ultra high accuracy option in both photo alignment and dense matching. In addition, the interpolation option is turned off in the generation of the DSM. Note that Metashape supports the processing of multi-view satellite images, but the results showed a very large elevation bias in this dataset. Therefore, in the experiment, only two of the views (FWD and BWD) are used.

Sat-MVSF. The inference settings of the proposed Sat-MVSF framework are as follows. In the pre-processing, selection of stereo pairs and bundle adjustment are not performed with this dataset. For the radiometric processing, τ is set to 2; the number of views is set to $n = 3$; a three-stage pyramid ($k = 3$) is applied; the intervals for the stages are $I_1 = (SR_{H1} - SR_{L1})/64$, $I_2 = 5$ m, and $I_3 = 2.5$ m, respectively; the numbers of hypothetical height planes are $N_1 = 64$, $N_2 = 32$, and $N_3 = 8$, respectively; and in the post-processing, $\psi = 1$ and $z = 2$. DSMs are generated after conversion from point cloud to grid.

Except for the MVS-based Adapted COLMAP and Sat-MVSF that directly process three-view images, the other methods work on stereo pairs with the requirement of pair-wise epipolar rectification.

4.2.2. Results

The results are listed in Table 1. It should be noted that some methods use interpolation to fill those invalid grids of a DSM as the corresponding image pixels are covered by clouds. Interpolation introduces errors in these regions, which can seriously affect the accuracy rating. Therefore, these few regions covered by clouds were masked out before the evaluation.

From Table 1, we can draw the following conclusions.

(1) The proposed Sat-MVSF framework significantly outperforms all the other methods, including both the software and open-source solutions, in terms of the two accuracy metrics of RMSE and MAE. Sat-MVSF achieves an RMSE of 3.65 m, whereas the best software solution obtains an RMSE of 7.9 m, and the best open-source solution of Adapted COLMAP obtains an RMSE of 4.7 m.

(2) For the two completeness metrics, the proposed framework obtains the best reconstruction accuracy in $PAG_{2.5m}$, at 6 points higher than the second-best Adapted COLMAP and 18 points higher than ArcGIS. In $PAG_{7.5m}$, the proposed framework achieves the second-best result, at 2.5 points less than the best method of CATALYST. However, it should be noted that the interpolation used in CATALYST leads to an improvement in $PAG_{7.5m}$.

(3) The proposed framework shows also good efficiency, surpassing most of the solutions, and is only slightly slower than CATALYST.

Overall, the proposed deep learning based Sat-MVSF framework achieves the best performance. An example is shown in Fig. 7, where invalid matches are marked in white. The results of ArcGIS and CATALYST appear to have a high degree of completeness, however, this is because interpolation is applied to these in fact invalid regions to obtain a better visualization, but such over-interpolation definitely results in a loss of accuracy, as the accuracy of these methods on the RMSE metric is poor in Table 1.

Table 1
Evaluation of the different solutions on the WHU-TLC test set. The bold figures rank first and the underlined figures rank second in a certain metric.

Method	RMSE (m)	MAE (m)	$PAG_{2.5m}$ (%)	$PAG_{7.5m}$ (%)	Time (min)
S2P	10.089	3.158	54.96	73.37	17.04
SDRDIS	15.012	4.496	47.58	73.57	9.41
Adapted COLMAP	4.714	2.168	58.78	76.80	77.45
ArcGIS	10.689	4.607	48.88	77.71	6.82
CATALYST	7.939	3.454	52.31	**82.52**	**3.80**
Metashape	13.047	2.693	56.59	75.46	24.51
Sat-MVSF (proposed)	**3.654**	**1.895**	**64.82**	80.05	5.87

4.3. Alternative learning modules

In addition to the RED structure we used in Sat-MVSF, other multiple-resolution inference modules, e.g., Cas-MVSNet (Gu et al., 2020) and UCS-Net (Cheng et al., 2020), can also be applied to processing the satellite MVS images. In Table 2, all the networks are trained using the training strategy described in Section 4.2. It is worth noting that the numbers of hypothetical height planes in UCS-Net are adaptive, and the UCS-Net parameters we used are consistent with the original paper (Cheng et al., 2020), except that RPC warping was used instead of homography warping.

Due to the high memory consumption of the 3D convolution used in Cas-MVSNet and UCS-Net, we could only perform the prediction on small patches of the WHU-TLC dataset. For a fair comparison, the images were cropped into patches with a size of 2048 × 1472 pixels. From Table 2, it can be observed that the performances of the three inference models are similar. RED-Net obtains a slightly higher accuracy, while Cas-MVSNet and UCS-Net are a little faster.

We also compared the results of using different patch sizes in RED-Net. The results are listed in Table 3. From Table 3, it can be observed that there is little difference between the accuracy with different patch sizes, but the efficiency shows a clear advantage when processing the larger patch size. Thus, compared with Cas-MVSNet and UCS-Net, RED-Net is more user-friendly for remote sensing images because it can process larger images with a lower memory occupation. Nevertheless, all of these deep learning based inference models can be used in the proposed framework, which shows the generality of the proposed framework. Samples of the results obtained from the different inference models, with almost the same effects, are shown in Fig. 8.

4.4. Performance on the MVS3D dataset

The MVS3D dataset (Bosch et al., 2016) is a multiple-view stereo benchmark for satellite images, which provides 50 WorldView-3 images and airborne LiDAR data for the ground truth. However, the RPC parameters are not calibrated, and the matched point clouds are not geometrically consistent with the ground truth. The GSD of the panchromatic image is about 0.3 m, and the acquisition time covers from November 2014 to January 2016, while the ground truth was collected in June 2016. As a result, there are huge scene differences between the stereo images and between the images and the ground truth, which indicates that the reconstruction of these scenes is an extremely challenging task, and is especially unfavorable for deep learning based methods. Moreover, MVS3D is much smaller than the WHU-TLC dataset, and lacks enough training samples for the deep learning based methods.

4.4.1. Implementation details

A. Self-Refinement Details.

As there are not enough training samples, we use the model pretrained on another open-source dataset with the self-refinement strategy. This is also a good chance to test the generalization ability of the deep learning based MVS model on satellite images. We train Sat-MVSNet with homography warping on the WHU-MVS aerial dataset (Liu and Ji, 2020) for 30 epochs to obtain a pretrained model. Then the pretrained model with RPC warping is used to predict the height maps of all the image pairs. The geometric consistency check ($\psi = 1$ and $z = 1$) is performed to obtain high-accuracy pseudo labels from the predicted height maps. Then, the pseudo labels are used to retrain Sat-MVSNet for only one epoch on the MVS3D dataset.

B. Inference Details.

For a fair comparison, the same pre-processing and post-processing operations are performed for all the methods. For the selection of the stereo image pairs, we set $\alpha = 5°$, $\beta = 45°$, and $\gamma = 40°$ to rank all the

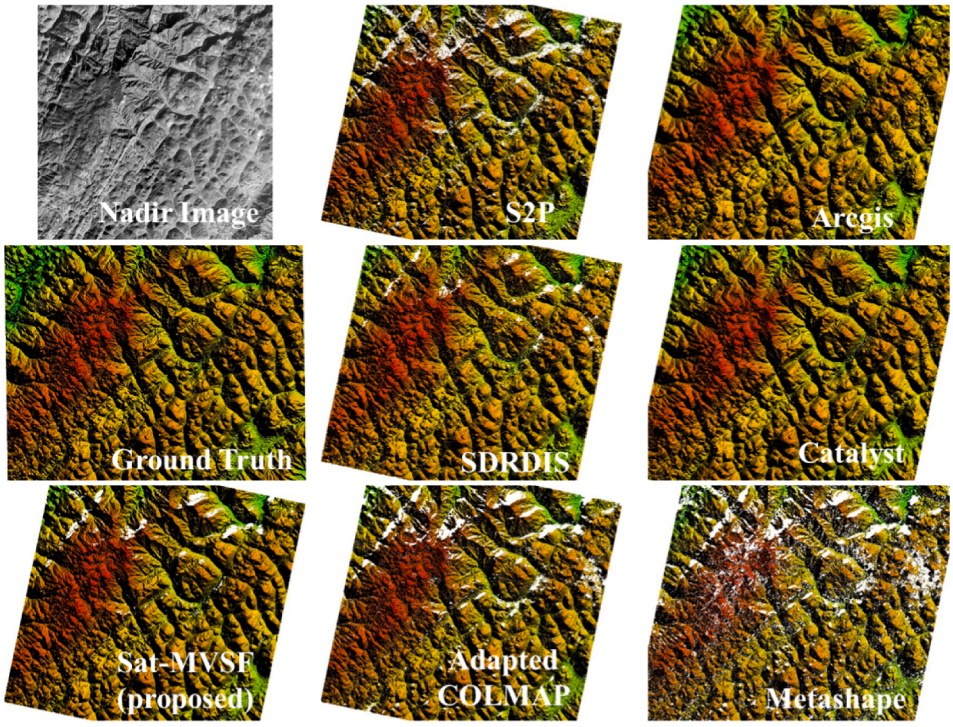

Fig. 7. Visualization examples of the results produced by the different solutions on the WHU-TLC dataset.

Table 2
Quantitative results of the different inference models on the WHU-TLC dataset.

Inference module	RMSE (m)	MAE (m)	$PAG_{2.5m}$ (%)	$PAG_{7.5m}$ (%)	Time (min)
Cas-MVSNet	**3.586**	1.995	63.25	79.44	**12.33**
UCS-Net	3.689	1.995	63.93	79.88	13.28
RED-Net	3.730	**1.900**	**64.79**	**80.06**	13.29

Table 3
Comparison with different patch sizes using RED-Net as the inference model.

Size	RMSE (m)	MAE (m)	$PAG_{2.5m}$ (%)	$PAG_{7.5m}$ (%)	Time (min)
2048 × 1472	3.730	1.900	64.79	80.06	13.29
5120 × 5120	3.654	1.895	64.82	80.05	5.87
Difference	−0.076	−0.005	+0.03	−0.01	−7.42

image pairs, and then select the top five pairs. For the bundle adjustment, the first image of each pair is selected as the reference, and no control points are considered. For the image cropping, patches with a size of 3072 × 3072 pixels are cropped using the officially provided cropping scheme (Bosch et al., 2016), to ensure that the area covered by the cropped patch is larger than the ground truth. The cropped offsets are absorbed into the RPC parameters to ensure that the form of the RPC warping remained unchanged. For the radiometric processing, a τ % linear stretch is used, where $\tau = 2$. In all the methods, the point cloud results are exported as input for the further processing. For CATALYST, where the point cloud results cannot be exported, we generate DSMs with a grid resolution approximately equal to that of the GSD and then convert them to point clouds. The officially provided tool (Bosch et al., 2016) is then used to achieve alignment between the point clouds and the ground truth DSM. Finally, DSMs with a resolution of 0.3 m are generated for the image pairs and fused according to Sections 3.7.2 and 3.7.3.

Except for the default settings, the other parameters are set as follows. In CATALYST (Catalyst, 2021), SGM is chosen for dense matching, and the scale of the epipolar resampling is set to 1 for higher matching quality. In Metashape (Agisoft, 2022), the ultra high accuracy option is selected for a better reconstruction. The bundle adjustment modules in SDRDIS (Sdrdis, 2016), JHU/APL example (the official solution for MVS3D dataset) (Bosch et al., 2016), and Adapted COLMAP (Zhang et al., 2019), are skipped. In the proposed Sat-MVSF framework, models respectively pretrained on the WHU-TLC, WHU-MVS datasets with additional self-refinement are used for the prediction. The view number is set to $n = 2$, and a three-stage pyramid is used. The search range is obtained by extending the SRTM elevation to a certain extent. The intervals of each stage are $I_1 = (SR_{H1} - SR_{L1})/64$, $I_2 = 0.6$ m, and $I_3 = 0.3$ m, and the numbers of hypothetical height planes are set to $N_1 = 64$, $N_2 = 32$, and $N_3 = 8$, respectively. For the geometric consistency check, the thresholds are set as $\psi = 1$ and $z = 1$.

4.4.2. Results

The evaluation results obtained on the MVS3D dataset are listed in Table 4. Several conclusions can be drawn from Table 4. Firstly, on such an extremely unfavorable dataset for learning methods, the model pretrained on the five-view WHU-MVS aerial dataset obtains the best median height error and RMSE scores averaged on all the sites. As to $PAG_{1.0m}$, the model pretrained on the WHU-MVS aerial dataset, plus self-refinement, obtains the best result. This is to say, the model trained on aerial data (five-view in this case) has great potential to be applied directly in the satellite dense matching task (two-view in this case), indicating that the knowledge a model learns from open-source MVS datasets is universal. This is of key importance for real-world applications.

Secondly, it can be observed that the model trained on the WHU-TLC dataset obtains worse results in two accuracy metrics, compared to the model trained on the WHU-MVS aerial dataset. The reason for this is that the WHU-MVS aerial dataset has five-view images and a GSD (0.1 m) close to that of the MVS3D dataset (0.3 m), but the WHU-TLC satellite dataset has three fixed views and a ground resolution of 2.5 m. This

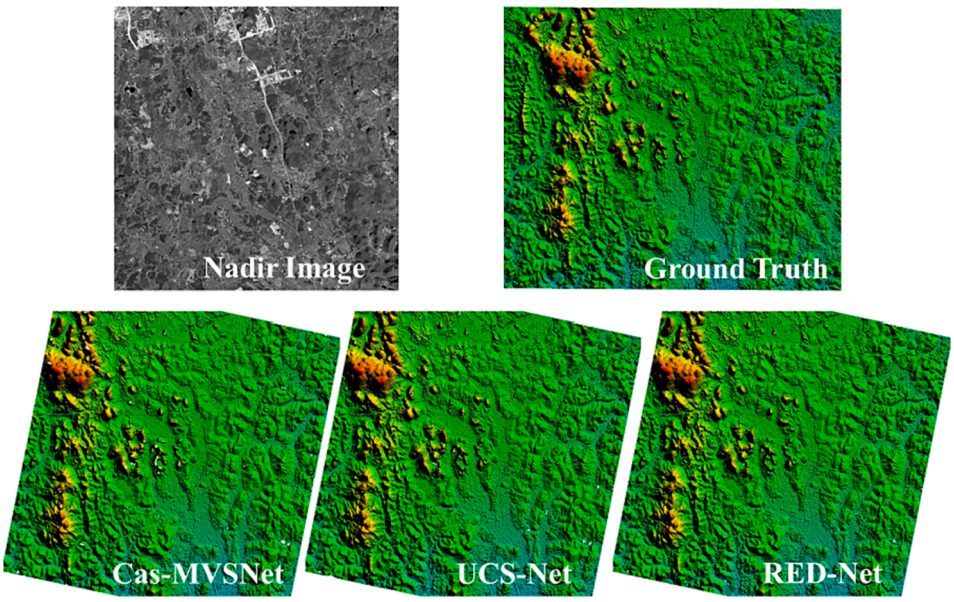

Fig. 8. Results of using different inference models when applied in the proposed framework.

Table 4
Evaluation results obtained on the MVS3D dataset. The bold figures rank first and the underlined figures rank second in each metric for each site. The self-refinement was applied on the model pretrained on the WHU-MVS aerial dataset.

Method		CATALYST	Metashape	S2P	SDRDIS	JHU/APL	Adapted COLMAP	Sat-MVSF (WHU-TLC)	Sat- MVSF (WHU-MVS)	Sat- MVSF (self-refinement)
Mean of all sites	$PAG_{1.0m}$(%)	58.915	56.73	59.49	56.67	55.19	50.38	55.90	51.09	**60.48**
	Median(m)	0.767	0.495	0.400	0.503	0.883	0.371	0.587	**0.368**	0.397
	RMSE(m)	4.323	3.464	4.778	4.166	4.896	8.397	3.867	**2.957**	3.242
Site1	$PAG_{1.0m}$(%)	72.31	67.61	**74.42**	69.82	68.09	63.21	68.47	51.44	72.51
	Median(m)	0.353	0.279	**0.235**	0.304	0.511	0.261	0.34	0.281	0.26
	RMSE(m)	2.913	2.495	2.416	2.772	3.156	3.468	2.83	**2.079**	2.492
Site2	$PAG_{1.0m}$(%)	64.57	65.65	**70.46**	63.82	61.91	55.64	63.78	69.15	70.07
	Median(m)	0.548	0.397	0.348	0.529	0.655	**0.264**	0.571	0.354	0.377
	RMSE(m)	2.037	1.872	1.836	2.038	2.182	5.464	2.045	**1.781**	1.885
Site3	$PAG_{1.0m}$(%)	**58.92**	54.22	55.06	57.65	54.05	46.7	54.47	50.34	58.47
	Median(m)	0.665	0.506	0.377	0.395	0.827	**0.321**	0.489	0.338	0.343
	RMSE(m)	4.311	3.898	3.874	3.912	4.581	9.596	3.795	**3.124**	3.340
Site4	$PAG_{1.0m}$(%)	**43.86**	38.68	40.16	41.37	41.94	28.83	39.64	24.4	42.03
	Median(m)	1.466	0.722	0.533	0.531	1.527	0.599	0.698	**0.389**	0.443
	RMSE(m)	10.319	7.214	12.873	7.844	11.749	19.138	7.8	**5.96**	7.085
Site5	$PAG_{1.0m}$(%)	65.58	66.01	**70.48**	63.65	61.73	64.22	63.59	70.04	70.28
	Median(m)	0.549	0.413	0.377	0.55	0.662	**0.35**	0.609	0.374	0.397
	RMSE(m)	2.127	1.821	1.772	2.055	2.24	2.777	2.142	**1.772**	1.887
Site6	$PAG_{1.0m}$(%)	63.32	62.85	**67.47**	61.17	57.44	60.88	60.65	66.23	66.99
	Median(m)	0.58	0.446	0.397	0.573	0.744	**0.375**	0.66	0.395	0.432
	RMSE(m)	2.611	2.105	2.102	2.519	2.63	3.29	2.743	**2.086**	2.27
Site7	$PAG_{1.0m}$(%)	42.26	44.66	44.38	40.91	42.16	37.02	41.8	33.41	**46.59**
	Median(m)	1.355	0.75	0.556	0.767	1.308	0.483	0.846	**0.451**	0.557
	RMSE(m)	6.192	4.651	6.353	4.652	8.162	13.192	5.869	**3.265**	3.681
Site8	$PAG_{1.0m}$(%)	**60.50**	54.14	53.50	54.95	54.19	46.51	54.82	43.71	56.88
	Median(m)	0.621	0.443	0.375	0.374	0.83	**0.316**	0.48	0.362	0.363
	RMSE(m)	4.077	3.657	6.996	7.539	4.467	10.254	3.713	3.592	**3.299**

indicates that, if the deep learning model was pretrained on a more suitable dataset closer to the target imagery, e.g., a future WorldView MVS dataset, the performance would be further improved.

Thirdly, the performance of the different methods varies at the different sites of the MVS3D dataset. The results of the different methods also vary in the three metrics. S2P achieves the highest completeness ($PAG_{1.0m}$) among the conventional methods, and approaches Sat-MVSF with self-refinement. Adapted COLMAP achieves the second-best median height error score but with the worst RMSE. Metashape achieves the best RMSE among the conventional methods, but this is clearly higher than that of Sat-MVSF pretrained on the WHU-MVS aerial dataset with self-refinement.

In conclusion, S2P can be considered as the best conventional method, and the refined Sat-MVSF framework surpasses the conventional methods on this challenging dataset which is especially difficult for deep learning based models.

Fig. 9 shows Site 1 of the MVS3D dataset, and the reconstructed surface obtained by the different methods. It can be seen that all the methods can restore the basic topographic profile of the area. The result of Sat-MVSF pretrained on the WHU-MVS aerial dataset contains a lot of voids. This is mainly caused by the significant differences in platforms, spectral characteristics, number of image views, intersection angles, resolution, etc. between the aerial training dataset and the target Worldview dataset. But when the self-training strategy is applied, semi-

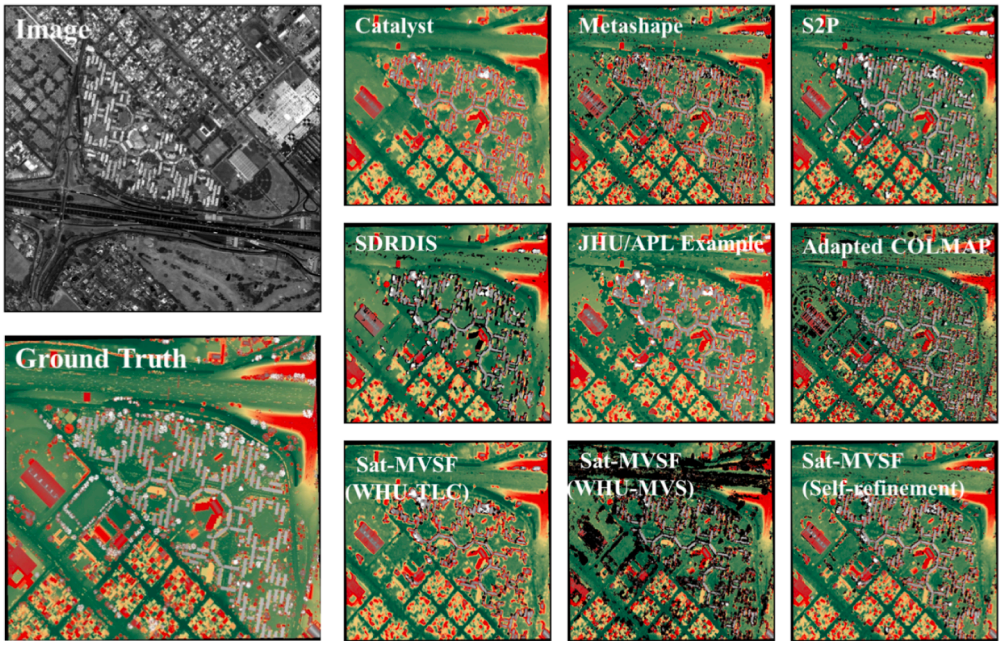

Fig. 9. Visualization of the results of the various methods in the first scene of the MVS3D dataset.

dense accurate labels can be generated to retrain the model to fill the gap between the two datasets. Finally, the proposed method achieves the best RMSE, and the completeness (PAG) is enhanced significantly with the help of the proposed self-refinement strategy. The refined Sat-MVSF method and S2P obtain the finest structures. CATALYST and JHU/APL appear to achieve the best $PAG_{1.0m}$, but in fact many of the color pixels are invalid in the calculation of the completeness score. S2P and the refined Sat-MVSF actually obtain the best completeness, according to Table 4.

We investigated the difference between models pretrained on the WHU-TLC and WHU-MVS datasets. As shown in Fig. 10a, in the building region, the predictions of the model pretrained on the WHU-TLC dataset are generally correct but not that fine. The model pretrained on the WHU-MVS aerial dataset performs better in the details. This proves our former conclusion that the WHU-TLC dataset lacks knowledge for higher-resolution reconstruction. Fig. 10b is the error map of refined Sat-MVSF pretrained on WHU-TLC. It can be seen that the boundaries of the buildings and small objects such as trees are the main error sources. Fig. 10c is the error map of the refined model pretrained on WHU-MVS, which has fewer errors than Fig. 10b. Nevertheless, due to the seasonal variations, there are differences between the given ground truth and the real ground truth in the vegetation areas.

4.4.3. *Different numbers of loops for the self-refinement strategy*

The applied self-refinement strategy does not require the involvement of any ground-truth labels, but is not costless. The self-refinement, involving the prediction, pseudo-label generation, and retraining, takes nearly 2 h for one loop.

We check whether we should perform more loops to obtain better results in the trade-off with efficiency. Based on the model pretrained on the WHU-MVS dataset, further experiments are performed on the MVS3D dataset. We increase the number of self-refinement loops and perform DSM production and accuracy evaluation in each loop. The experimental results are shown in Fig. 11, for which the metrics are the same as those described in Section 4.1.

According to Fig. 11, The first loop brings a substantial improvement in completeness, whereas the RMSE and median error scores correspondingly go worse. Nevertheless, the first loop is necessary because PAG is a more comprehensive indicator considering both accuracy and completeness. After the first loop, the completeness is slightly increased in the following loops, accompanied by a slight loss of accuracy. This experiment illustrates that the self-refinement strategy only needs to be

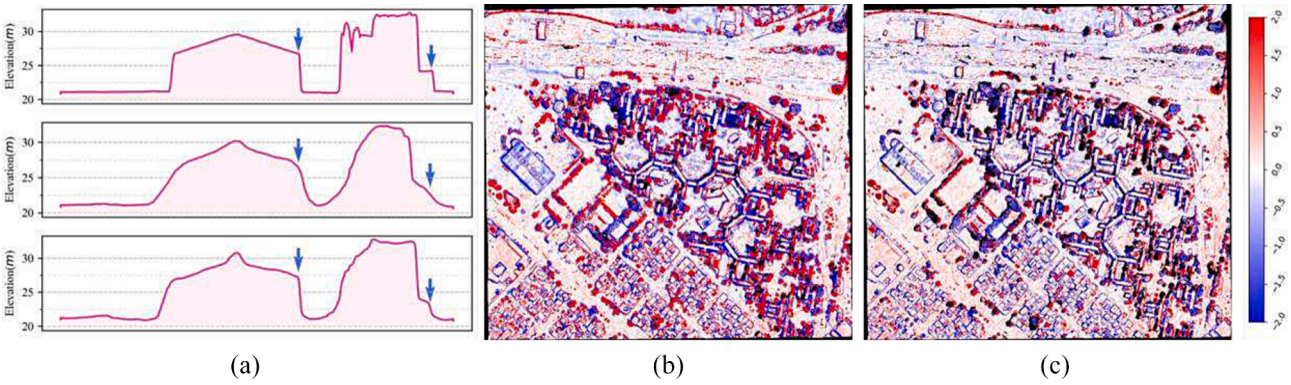

Fig. 10. (a) Profiles of the ground truth (top) of a building region, and the results obtained by the model pretrained on the WHU-TLC (middle) and WHU-MVS datasets (bottom). (b) Error map of the DSM produced by the refined Sat-MVSF pretrained on the WHU-TLC dataset. (c) Error map of the DSM produced by the refined Sat-MVSF pretrained on the WHU-MVS dataset. (b) and (c) share the same color bar shown on the right.

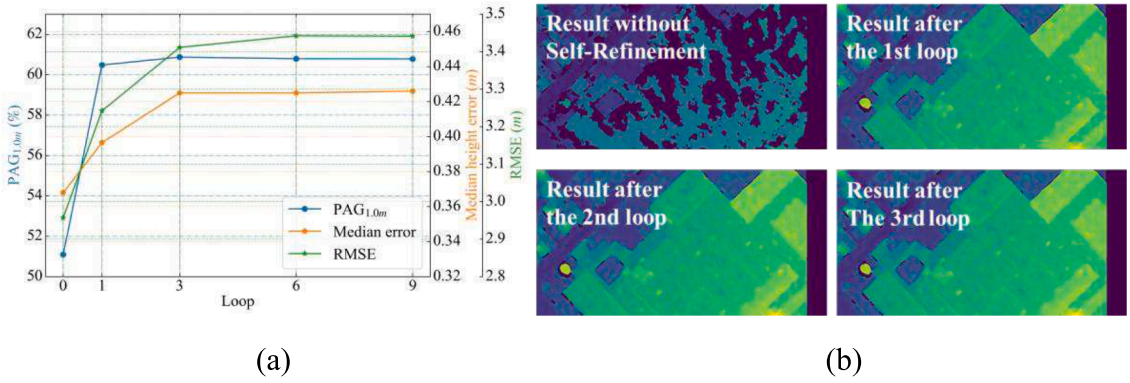

Fig. 11. (a) Changes in $PAG_{1.0m}$, median height error, and RMSE as the number of self-refinement loops increases. Note that a higher completeness score means better reconstruction, while a lower RMSE and median error score means better accuracy. (b) The predicted height map results obtained in the different loops.

performed for one or two loops to achieve the best results. This is the ideal solution for practical applications, with a lower cost and considerable payoff, and, more importantly, no labels of target regions are required.

4.4.4. Test on true MVS mode

We have evaluated the proposed Sat-MVSF and the other methods in a pair-wise mode on the MVS3D dataset. Among these methods, only Sat-MVSF and Adapted-COLMAP can work on a true MVS mode, i.e., using all available multi-view images as input instead of a stereo pair. In this section, we compare the performance of the two methods on the MVS mode. We use the same parameter settings and self-refined network weights described in 4.4.1. The results are shown in Table 5.

According to Table 5, using the multi-view input brings a 5.0 % PAG improvement for Adapted-COLMAP and a 2.5 % decrease for Sat-MVSF, a 48.5 % RMSE improvement for Adapted-COLMAP and an 8.0 % improvement for Sat-MVSF, compared to the use of stereo pairs. Nevertheless, Sat-MVSF still outperforms Adapted-COLMAP both in PAG and RMSE in the MVS mode. The huge improvement of COLMAP, as illustrated in Fig. 11, is because it employs a dedicated pixel-by-pixel view selection algorithm (Schönberger et al., 2016), which is able to find, for each pixel, the optimal multi-view image set suitable for matching, avoid some inconsistent information between images caused by noise, occlusion, etc., and maximize the benefits of multi-view information injection. Sat-MVSF does not have such a pixel-level view selection mechanism but treats multi-view information equally, which results in a slight decrease in the PAG metric although many voids have been filled as shown in Fig. 12. Nevertheless, such view selection mechanism can be considered to implement in Sat-MVSF in the future.

Overall, our proposed learning-based Sat-MVSF works well not only in the pair-wise mode, but also in the MVS mode, showing its high applicability.

4.5. More examples

We also apply Sat-MVSF to the Sainte-Maxime and Hong Kong datasets from the ZY3-01 satellite provided by the ISPRS (ISPRS, 2018), as further qualitative observations. The GSDs of the ZY3-01 nadir-view, front-view, and back-view images are 2.1 m, 3.5 m, and 3.5 m. This is close to the image settings in the WHU-TLC dataset. For the Sainte-Maxime dataset, the control points provided are used for the bundle adjustment. For the Hong Kong dataset, due to the lack of control points, the nadir-view image is selected as the reference image for the bundle adjustment. The other parameters are the same as in Section 4.2.1 B. The model is trained on the WHU-TLC dataset and no self-refinement was performed.

Furthermore, we use the SRTM DEM Version 4 (Jarvis et al., 2008) to fill some voids in the DSM generated by the framework, for better visualization. The final results of the proposed Sat-MVSF are shown in Fig. 13 where water has been artificially masked off.

Table 5
Results (mean of all sites) of true MVS methods on the MVS3D dataset with pair-wise and MVS mode, the latter means all available multi-view images are used as input.

Method	Mode	$PAG_{1.0m}$ (%)	Median(m)	RMSE(m)
Adapted-COLMAP	Pair-wise	50.38	0.371	8.397
	MVS	55.35	0.404	4.320
Sat-MVSF	Pair-wise	60.48	0.397	3.242
	MVS	58.01	0.501	2.982

5. Discussion

Deep learning methods have freed us from the limitations of hand-crafted features by learning deep features and knowledge directly from data. In this study, the proposed deep learning model has exhibited a great performance and showed great potential in the field of 3D reconstruction from MVS satellite images, despite the very unfavorable situation of high-quality training samples being extremely lacking.

We have access to only one dataset, i.e., the WHU-TLC dataset that we created recently (Gao et al., 2021), that is appropriate for the training and testing of deep learning based MVS methods. In this dataset, the images are all collected from the same TLC sensor. The models pretrained on this dataset at a lower resolution (2.5 m) do not work very well in the reconstruction of higher-resolution images (MVS3D, 0.3 m), as expected. Instead, the model pretrained on the aerial MVS dataset (with 0.1 m GSD) obtains much better results. With the support of the self-refinement strategy, the pre-trained model beats the conventional methods without the requirement of target training samples, showing its high generalization ability. Nevertheless, the lack of sufficient training datasets is the biggest obstacle to the development of deep learning methods in the field of satellite 3D reconstruction. Richer satellite MVS datasets covering a variety of mainstream sensors, different resolutions and view angles, and larger regions with varied land-cover types, will definitely boost the development of the deep learning based methods.

6. Conclusion

In this paper, we have proposed a general and complete deep learning based framework for 3D surface reconstruction from MVS satellite images, which is named Sat-MVSF. The capability and potential of the deep learning approach has been demonstrated by a full comparison with both commercial software and open-source solutions. The experiments show that the proposed method achieves the best result when the inferred images are captured from the same sensor as the training dataset. Furthermore, the proposed framework achieves a slightly better

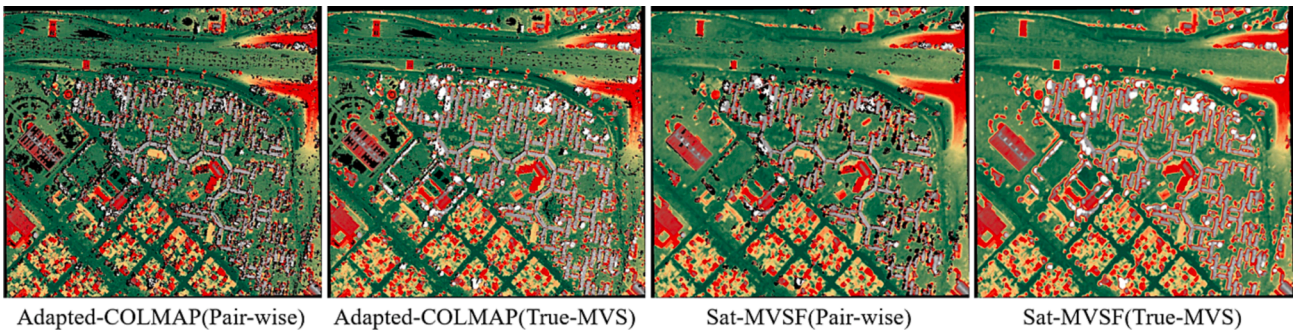

Fig. 12. Visualization of the results of true MVS methods with respective pair-wise and MVS mode on the MVSD dataset.

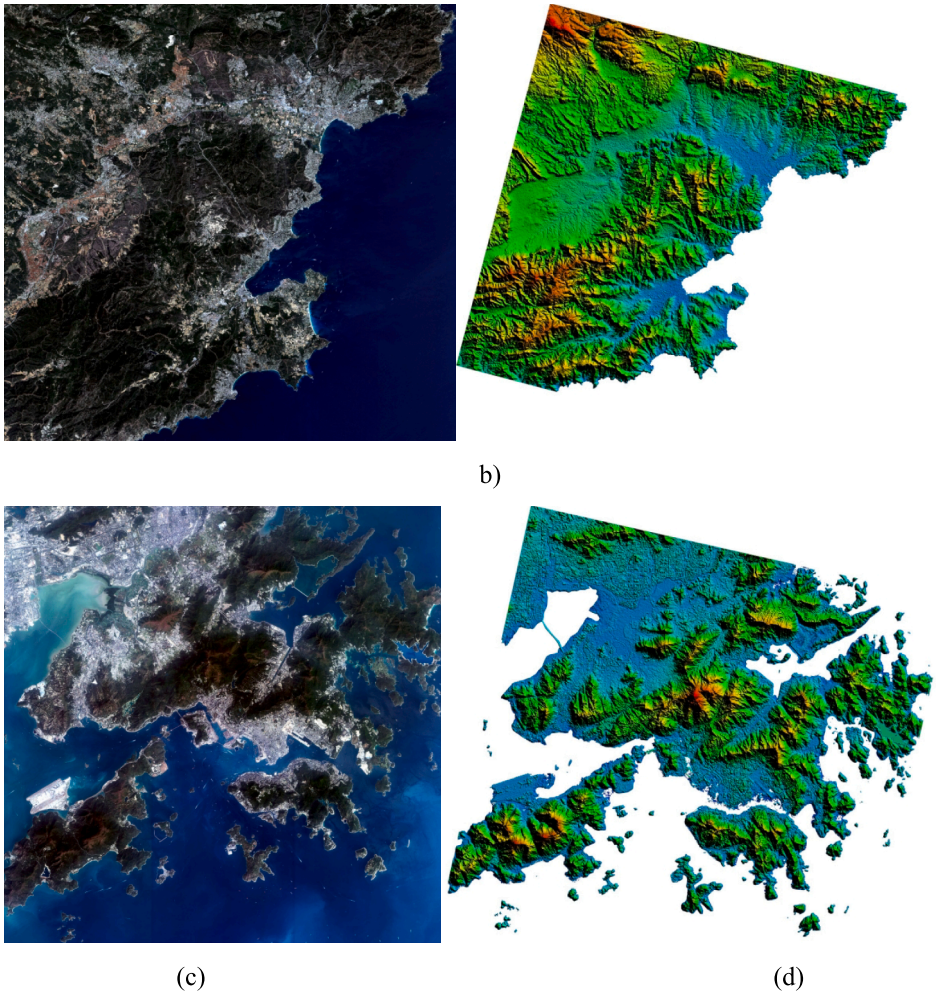

Fig. 13. The results produced by the proposed Sat-MVSF: (a) and (b) are the true color image and DSM product for the Sainte-Maxime dataset, respectively; and (c) and (d) are those for the Hong Kong dataset.

performance than the existing methods on test images that are very different from the training data. We also point out that the lack of richer datasets is the major obstacle to the development of deep learning in satellite MVS reconstruction. As an early exploration, we hope that this study will promote both the study of deep learning based satellite MVS methods and the creation of versatile open-source satellite MVS datasets.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (grant No. 42171430) and the State Key Program of the National Natural Science Foundation of China (grant No. 42030102). We thank the anonymous reviewers for their professional comments.

References

Agisoft, 2022. Agisoft Metashape [WWW Document]. URL https://www.agisoft.com/.

Ahn, C.-H., Cho, S.-I., Jeon, J.C., 2001. Ortho-rectification software applicable for IKONOS high resolution images: GeoPixel-Ortho, in: IGARSS 2001. Scanning the Present and Resolving the Future. Proceedings. IEEE 2001 International Geoscience and Remote Sensing Symposium (Cat. No. 01CH37217). IEEE, pp. 555–557.

ArcGIS, 2022. Ortho mapping with mosaic datasets—Help | ArcGIS Desktop [WWW Document]. URL https://desktop.arcgis.com/en/arcmap/10.5/manage-data/raster-and-images/ortho-mapping-overview.htm.

Bleyer, M., Rhemann, C., Rother, C., 2011. Patchmatch stereo-stereo matching with slanted support windows., in: Bmvc. pp. 1–11.

Bosch, M., Kurtz, Z., Hagstrom, S., Brown, M., 2016. A multiple view stereo benchmark for satellite imagery, in: 2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). IEEE, pp. 1–9.

Bosch, M., Foster, K., Christie, G., Wang, S., Hager, G.D., Brown, M., 2019. Semantic stereo for incidental satellite images, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 1524–1532.

Bradski, G., 2000. The openCV library. Dr. Dobb's Journal: Software Tools for the Professional Programmer 25, 120–123.

Catalyst, 2021. Catalyst Professional – CATALYST.Earth [WWW Document]. URL https://catalyst.earth/products/catalyst-pro/.

Cheng, S., Xu, Z., Zhu, S., Li, Z., Li, L.E., Ramamoorthi, R., Su, H., 2020. Deep stereo using adaptive thin volume representation with uncertainty awareness, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2524–2534.

Collins, R.T., 1996. A space-sweep approach to true multi-image matching, in: Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Ieee, pp. 358–363.

de Franchis, C., Meinhardt-Llopis, E., Michel, J., Morel, J.-M., Facciolo, G., 2014. An automatic and modular stereo pipeline for pushbroom images. ISPRS Ann. Photogr., Remote Sens. Spatial Information Sci. 49–56.

Ding, Y., Yuan, W., Zhu, Q., Zhang, H., Liu, Xiangyue, Wang, Y., Liu, Xiao, 2022. Transmvsnet: Global context-aware multi-view stereo network with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8585–8594.

Facciolo, G., de Franchis, C., Meinhardt, E., 2015. MGM: A significantly more global matching for stereovision, in: BMVC 2015.

Facciolo, G., de Franchis, C., Meinhardt-Llopis, E., 2017. Automatic 3D reconstruction from multi-date satellite images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 57–66.

Fan, D., Song, J., Rong, L., Yongsheng, Z., 2007. Automatic DSM Generation form Aerial Three Line Array ADS40 Digital Images, in: 2007 8th International Conference on Electronic Measurement and Instruments. IEEE, pp. 2–870.

Fraser, C.S., Hanley, H.B., 2003. Bias compensation in rational functions for IKONOS satellite imagery. Photogramm. Eng. Remote Sens. 69, 53–57.

Gao, J., Liu, J., Ji, S., 2021. Rational polynomial camera model warping for deep learning based satellite multi-view stereo matching, in: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6148–6157.

Grodecki, J., 2001. IKONOS stereo feature extraction-RPC approach, in: ASPRS Annual Conference St. Louis.

Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P., 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2495–2504.

Helava, U. v, 1988. Object-space least-squares correlation, in: (ACSM and American Society for Photogrammety and Remote Sensing, Annual Convention, Saint Louis, MO, Mar. 14-18, 1988) Photogrammetric Engineering and Remote Sensing,. pp. 711–714.

Hirschmuller, H., 2005. Accurate and efficient stereo processing by semi-global matching and mutual information, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, pp. 807–814.

ISPRS, 2018. ISPRS Data sets: ZY-3 [WWW Document]. URL https://www.isprs.org/data/zy-3/Default-HongKong-StMaxime.aspx.

Jarvis, A., Reuter, H.I., Nelson, A., Guevara, E., 2008. Hole-filled SRTM for the globe Version 4. available from the CGIAR-CSI SRTM 90m Database (http://srtm. csi. cgiar. org) 15, 5.

Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H., 2014. Large scale multi-view stereopsis evaluation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 406–413.

Knapitsch, A., Park, J., Zhou, Q.-Y., Koltun, V., 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Trans. Graph. (ToG) 36, 1–13.

Krauß, T., d'Angelo, P., Schneider, M., Gstaiger, V., 2013. The fully automatic optical processing system CATENA at DLR. ISPRS Hannover Workshop. 177–181.

Kuschk, G., d'Angelo, P., Qin, R., Poli, D., Reinartz, P., Cremers, D., 2014. DSM accuracy evaluation for the ISPRS commission I image matching benchmark. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. 40, 195–200.

Leys, C., Ley, C., Klein, O., Bernard, P., Licata, L., 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. J. Exp. Soc. Psychol. 49, 764–766.

Liu, J., Ji, S., 2020. A novel recurrent encoder-decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6050–6059.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60, 91–110.

Marí, R., de Franchis, C., Meinhardt-Llopis, E., Facciolo, G., 2019. To bundle adjust or not: A comparison of relative geolocation correction strategies for satellite multi-view stereo, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 2188–2196.

Meng, H., Liu, Y., Zhang, J., Gong, H., 2007. Positional accuracy in RPC point determination based on high-resolution imagery, in: Geoinformatics 2007: Remotely Sensed Data and Information. SPIE, pp. 1439–1449.

Moratto, Z.M., Broxton, M.J., Beyer, R.A., Lundy, M., Husmann, K., 2010. Ames Stereo Pipeline, NASA's open source automated stereogrammetry software, in: Lunar and Planetary Science Conference. p. 2364.

Ozcanli, O.C., Dong, Y., Mundy, J.L., Webb, H., Hammoud, R., Victor, T., 2014. Automatic geo-location correction of satellite imagery, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 307–314.

Paderes, F., 1989. Batch and on-line evaluation of stereo SPOT imagery, in: 1989 ASPRS/ACSM Annual Convention, Baltimore, MD. pp. 31–40.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., 2019. Pytorch: an imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32, 8026–8037.

Qin, R., 2016. Rpc stereo processor (rsp)–a software package for digital surface model and orthophoto generation from satellite stereo imagery. ISPRS Ann. Photogram., Remote Sens. Spatial Inform. Sci. 3, 77–82.

Rupnik, E., Daakir, M., Pierrot Deseilligny, M., 2017. MicMac–a free, open-source solution for photogrammetry. Open Geospat. Data, Softw. Standards 2, 1–9.

Rupnik, E., Pierrot-Deseilligny, M., Delorme, A., 2018. 3D reconstruction from multi-view VHR-satellite images in MicMac. ISPRS J. Photogramm. Remote Sens. 139, 201–211.

Scharstein, D., Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vis. 47, 7–42.

Schönberger, Johannes L, Frahm, J.-M., 2016. Structure-from-motion revisited, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4104–4113.

Schönberger, Johannes L, Zheng, E., Frahm, J.-M., Pollefeys, M., 2016. Pixelwise view selection for unstructured multi-view stereo, in: European Conference on Computer Vision. Springer, pp. 501–518.

Schönberger, Johannes L., Zheng, E., Frahm, J.M., Pollefeys, M., 2016. Pixelwise view selection for unstructured multi-view stereo, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-319-46487-9_31.

Schops, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A., 2017. A multi-view stereo benchmark with high-resolution images and multi-camera videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3260–3269.

Sdrdis, 2016. sdrdis/iarpa: My IARPA Contest submission [WWW Document]. URL https://github.com/sdrdis/iarpa.

Tachikawa, T., Kaku, M., Iwasaki, A., Gesch, D.B., Oimoen, M.J., Zhang, Z., Danielson, J. J., Krieger, T., Curtis, B., Haase, J., 2011. ASTER global digital elevation model version 2-summary of validation results. NASA.

Tao, C.V., Hu, Y., 2001. A comprehensive study of the rational function model for photogrammetric processing. Photogramm. Eng. Remote Sens. 67, 1347–1358.

Tao, C.V., Hu, Y., Jiang, W., 2004. Photogrammetric exploitation of IKONOS imagery for mapping applications. Int. J. Remote Sens. 25, 2833–2853.

Toutin, T., 2001. Geometric processing of IKONOS Geo images with DEM, in: ISPRS Joint Workshop High Resolution from Space. Citeseer, pp. 19–21.

Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M., 2021. Patchmatchnet: Learned multi-view patchmatch stereo, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14194–14203.

Wang, M., Hu, F., Li, J., 2011. Epipolar resampling of linear pushbroom satellite imagery by a new epipolarity model. ISPRS J. Photogramm. Remote Sens. 66, 347–355.

Wei, Z., Zhu, Q., Min, C., Chen, Y., Wang, G., 2021. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6187–6196.

Xiong, Z., Zhang, Y., 2010. Bundle adjustment with rational polynomial camera models based on generic method. IEEE Trans. Geosci. Remote Sens. 49, 190–202.

Xu, Q., Su, W., Qi, Y., Tao, W., Pollefeys, M., 2022. Learning inverse depth regression for pixelwise visibility-aware multi-view stereo networks. Int. J. Comput. Vis. 130, 2040–2059.

Yan, J., Wei, Z., Yi, H., Ding, M., Zhang, R., Chen, Y., Wang, G., Tai, Y.-W., 2020. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. Eur. Conf. Comput. Vision. Springer 674–689.

Yang, J., Mao, W., Alvarez, J.M., Liu, M., 2020. Cost volume pyramid based depth inference for multi-view stereo, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4877–4886.

Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L., 2018. Mvsnet: Depth inference for unstructured multi-view stereo, in: Proceedings of the European Conference on Computer Vision (ECCV). pp. 767–783.

Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L., 2019. Recurrent mvsnet for high-resolution multi-view stereo depth inference, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5525–5534.

Yu, A., Guo, W., Liu, B., Chen, X., Wang, X., Cao, X., Jiang, B., 2021. Attention aware cost volume pyramid based multi-view stereo network for 3d reconstruction. ISPRS J. Photogramm. Remote Sens. 175, 448–460.

Yu, D., Ji, S., Liu, J., Wei, S., 2021. Automatic 3D building reconstruction from multi-view aerial images with deep learning. ISPRS J. Photogramm. Remote Sens. 171, 155–170.

Zhang, K., Snavely, N., Sun, J., 2019. Leveraging Vision Reconstruction Pipelines for Satellite Imagery, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 2139–2148. https://doi.org/10.1109/ICCVW.2019.00269.

Zhang, X., Hu, Y., Wang, H., Cao, X., Zhang, B., 2021. Long-range attention network for multi-view stereo, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3782–3791.
Zhang, L., Gruen, A., 2006. Multi-image matching for DSM generation from IKONOS imagery. ISPRS J. Photogramm. Remote Sens. 60, 195–211.
Zhang, J., Li, S., Luo, Z., Fang, T., Yao, Y., 2022. Vis-MVSNet: visibility-aware multi-view stereo network. Int. J. Comput. Vis. 1–16.
Zheng, Z., Yang, Y., 2021. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. Int. J. Comput. Vis. 129, 1106–1120.