DOI: 10.1111/phor.12471

ORIGINAL ARTICLE

Digital surface model generation from high-resolution satellite stereos based on hybrid feature fusion network

Zhi Zheng[1,2] | Yi Wan[1] | Yongjun Zhang[1] | Zhonghua Hu[1] | Dong Wei[1] | Yongxiang Yao[1] | Chenming Zhu[3] | Kun Yang[4] | Rang Xiao[5]

[1]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China

[2]Department of Geography and Resource Management, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China

[3]The First Institute of Photogrammetry and Remote Sensing, Ministry of Natural Resources, Xi'an, China

[4]Guizhou Basic Geographic Information Center, Guiyang, China

[5]Guizhou Tuzhi Information Technology, Guiyang, China

Correspondence
Yi Wan and Yongjun Zhang, School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.
Email: yi.wan@whu.edu.cn; zhangyj@whu.edu.cn

Abstract

Recent studies have demonstrated that deep learning-based stereo matching methods (DLSMs) can far exceed conventional ones on most benchmark datasets by both improving visual performance and decreasing the mismatching rate. However, applying DLSMs on high-resolution satellite stereos with broad image coverage and wide terrain variety is still challenging. First, the broad coverage of satellite stereos brings a wide disparity range, while DLSMs are limited to a narrow disparity range in most cases, resulting in incorrect disparity estimation in areas with contradictory disparity ranges. Second, high-resolution satellite stereos always comprise various terrain types, which is more complicated than carefully prepared datasets. Thus, the performance of DLSMs on satellite stereos is unstable, especially for intractable regions such as texture-less and occluded regions. Third, generating DSMs requires occlusion-aware disparity maps, while traditional occlusion detection methods are not always applicable for DLSMs with continuous disparity. To tackle these problems, this paper proposes a novel DLSM-based DSM generation workflow. The workflow comprises three steps: pre-processing, disparity estimation and post-processing. The pre-processing step introduces low-resolution terrain to shift unmatched disparity ranges into a fixed scope and crops satellite stereos to regular patches. The disparity estimation step proposes a hybrid feature fusion network (HF^2Net) to improve the matching performance. In detail, HF^2Net designs a cross-scale feature extractor (CSF) and a multi-scale cost filter. The feature extractor differentiates structural-context features in complex scenes and thus enhances HF^2Net's robustness to satellite stereos, especially on intractable regions. The cost filter filters out most matching errors to ensure accurate disparity estimation. The post-processing step generates initial DSM patches with estimated disparity maps and then refines them for the final large-scale DSMs. Primary experiments on the public US3D dataset showed better accuracy than state-of-the-art methods, indicating HF^2Net's superiority. We then created a self-made Gaofen-7 dataset to train HF^2Net and conducted DSM generation experiments on two Gaofen-7 stereos to further demonstrate the effectiveness and practical capability of the proposed workflow.

KEYWORDS
deep learning-based stereo matching (DLSM), DSM generation, hybrid feature fusion network (HF^2Net), satellite stereos

INTRODUCTION

The Digital Surface Model (DSM), as one of the most fundamental geographical products, has been widely used in various applications, such as 3D city modelling, 3D land-use management and disaster monitoring (Gruen et al., 2013; Han et al., 2020; Liu et al., 2023; Lv et al., 2022; Lv, Zhong, Wang, You, & Falco, 2023; Lv, Zhong, Wang, You, & Shi, 2023; Zhao et al., 2022). Satellite stereos, due to their characteristics of flexible acquisition and low cost, have been the dominant data source for generating city- or country-level DSMs (Bosch et al., 2016; Gao et al., 2021; Huang et al., 2016; Kendall et al., 2017; Leotta et al., 2019; Lv, Zhong, Wang, You, & Falco, 2023; Zhang, Cui, et al., 2022). Thus, satellite image stereo matching (SISM) continues to be a hot research topic in recent years (Huang & Qin, 2020; Michel et al., 2020; Qin, 2019a; Zhang et al., 2017, 2019). Conventional dense matching algorithms generally calculate the matching cost through manually designed feature descriptors and then estimate disparity values with designed matching functions (Scharstein & Szeliski, 2002). However, limited by the insufficient description capability, conventional algorithms always suffer from serious mismatching problems in intractable regions, including texture-less and repetitive regions (i.e., farmland), heavily occluded regions (i.e., dense buildings) and other terrains (Facciolo et al., 2017; Huang et al., 2018; Qin, 2019b). Recent works show that the mismatching rates of stereo matching can be significantly decreased by using deep-learning technical as solvers (Gao et al., 2021; Ji et al., 2019; Shen et al., 2020). However, though deep learning-based stereo matching methods (DLSMs) have flourished in recent years, applying DLSMs on high-resolution satellite stereos with broad image coverage and wide terrain variety is still challenging (Bosch et al., 2016; Chang & Chen, 2018; Shen et al., 2021; Xu et al., 2022).

The complicated image content of satellite stereos is one of the main obstacles. Unlike carefully prepared benchmark datasets, satellite images comprise various terrain types, such as texture-less areas, severe occluded regions and steep mountains. Since it is unrealistic to establish datasets with complete topographical categories for DLSMs, the complicated image content in satellite stereos raises higher accuracy and robustness demand for DLSMs (Cournet et al., 2020; Schops et al., 2017). The dramatic disparity range of satellite stereos is another hurdle. Due to the significant viewing difference and drastic elevation changes in broad coverage scenes, the disparity values in satellite stereos vary widely and change from negative to positive. However, most DLSMs can only estimate correct disparity values within a narrow scope due to the memory restrictions of graphics processing unit (GPU) (Gao et al., 2023; He et al., 2022). In addition, how to obtain occlusion-aware disparity values for DSMs generation must be investigated. Conventional workflows use left–right consistency check methods to filter out matching errors and occluded pixels (Huang et al., 2018; Zhang et al., 2019). However, these methods are not always applicable to DLSMs with continuous disparity estimation.

To address the problems above-mentioned, we propose a novel DLSM-based DSM generation workflow, including pre-processing, disparity estimation and post-processing. The pre-processing step shifts unmatched disparity ranges into a fixed scope and crops satellite stereos to regular patches. The disparity estimation step then predicts a complete disparity map for each patch with the designed hybrid feature fusion network (HF^2Net). At last, the post-processing step generates initial DSM patches with estimated disparity maps and then refines them to obtain the final large-scale DSMs. The primary contributions of this paper can be summed up as follows:

- A novel DLSM named HF^2Net for SISM. In detail, HF^2Net designs a CSF for structural-context feature differentiation and a multi-scale cost filtering module to filter out most matching errors. Thus, HF^2Net ensures accurate disparity estimation in complex scenes and thus enables robust matching on satellite stereos.
- A complete DLSM-based workflow for large-scale DSM generation. In the workflow, the pre-processing step provides regular stereo patches, the disparity estimation step predicts accurate disparity values and the post-processing step obtains DSMs and refines the results. The obstacles of applying DLSMs to satellite stereos are gradually handled throughout the workflow.
- Systematic performance evaluation for HF^2Net and the proposed workflow. Experiments on the public US3D dataset and Gaofen-7 stereos showed the superiority of HF^2Net and the application capability of the proposed workflow. For example, the proposed HF^2Net reduced the mismatching rate by approximately 15% in texture-less areas compared with the selected conventional method.

The rest of this paper is organized as follows. Section 2 overviews the relevant literature. Section 3 describes the proposed HF^2Net and the built workflow for large-scale DSM generation in detail. Section 4 depicts the comprehensive investigation of the proposed HF^2Net and discusses the superiority and shortcomings of the proposed workflow. Section 5 concludes with the findings and provides recommendations for future work.

RELATED WORK

Conventional stereo matching methods include four steps: matching cost calculation, cost aggregation, disparity estimation and disparity refinement (Scharstein & Szeliski, 2002). Although significant progress has been made in decades for conventional methods (Facciolo et al., 2017; Hou et al., 2018; Huang et al., 2016; Youssefi et al., 2020; Zhang et al., 2017; Zhang, Zou, et al., 2022), they still face the problems of high mismatching rates and poor performance in intractable regions due to the limitation of manually designed descriptors and matching functions.

With the rapid development of deep-learning technology, researchers formulated the stereo matching task as a supervised task and used deep neural networks as solvers. Early works used artificial neurons to calculate the matching cost between stereo images, and the remaining steps were still following the process of conventional methods (Zbontar & LeCun, 2016; Zhang & Wah, 2017). Although higher matching accuracy has been achieved, these methods are no essential difference from conventional methods. That is, there still exists error accumulation among different steps. The milestone achievement of DLSMs was GCNet (Kendall et al., 2017), where a novel convolutional neural network was designed to formulate an end-to-end stereo matching procedure. Thereafter, DLSMs flourished in the computer vision and photogrammetry fields. Chang and Chen (2018) proposed PSMNet, which utilized spatial pyramid pooling to exploit multi-scale features and stacking multiple hourglass modules for cost filtering. Guo et al. (2019) introduced GwcNet, where the cost volume is built by group-wise correlation to measure the image similarities. Rao et al. (2020) joined semantics and geometry to estimate the disparity and used a non-local context attention module for cost volume regularization. Xu et al. (2022) proposed an attention concatenation volume (ACV) to ease the burden of cost volume while maintaining state-of-the-art matching accuracy.

Due to the lack of suitable training datasets, the development of DLSMs for satellite stereos is lagged behind its development for natural images. Recent works for SISM first compared the performance difference between conventional methods and DLSMs (Albanwan & Qin, 2022). Chen et al. (2019) compared the cost function of Census (Hirschmuller, 2007) and fast-CNN (Zbontar & LeCun, 2015) to evaluate the performance influence of both low- and high-level features on DLSM methods. Ji et al. (2019) compared the performance of the traditional SGM (Hirschmuller, 2007) with several DLSMs. Other researchers applied state-of-the-art methods to SISM. For example, Qin et al. (2019) applied PSMNet for the disparity estimation of World-View images and obtained convincing results. There are also some creative networks for SISM. For example, Rao et al. (2020) proposed BGANet, which joined semantic segmentation and disparity estimation tasks in a unified framework. Although some methods were proposed to deal with satellite stereos, they mainly focused on the accuracy of the estimated disparity rather than the final DSMs. Thus, the testing geographical scenes in their research were relatively simple, and the disparity was fixed to a small scope (Tao et al., 2020), which is not extensive enough for city- or country-level DSM generation.

METHODOLOGY

This paper establishes a practical DLSMs-based workflow for large-scale DSM generation, which comprises pre-processing, disparity estimation and post-processing steps. In the disparity estimation step, this paper contributes a novel network named HF^{2}Net to the remote sensing community, aiming at enhancing DLSMs' performance on complicated satellite stereos. Since disparity estimation is the most essential step in the workflow, we first detail the architecture of the proposed HF^{2}Net in Section 3.1. We then systemically introduce the workflow in Section 3.2.

Hybrid feature fusion network (HF^{2}Net)

Figure 1 shows the schematic architecture of HF^{2}Net, including four main parts: feature extraction, cost volume construction, multi-scale cost filtering and disparity regression. The architecture of HF^{2}Net is inherited from PSMNet (Chang & Chen, 2018), where the main modifications are the CSF and the multi-scale cost filtering module. The CSF takes two asymmetric branches to differentiate structural-context features and thus enhance HF^{2}Net's matching ability on satellite stereos, especially for intractable regions. The multi-scale cost filtering module filters out most matching errors to ensure accurate disparity estimation.

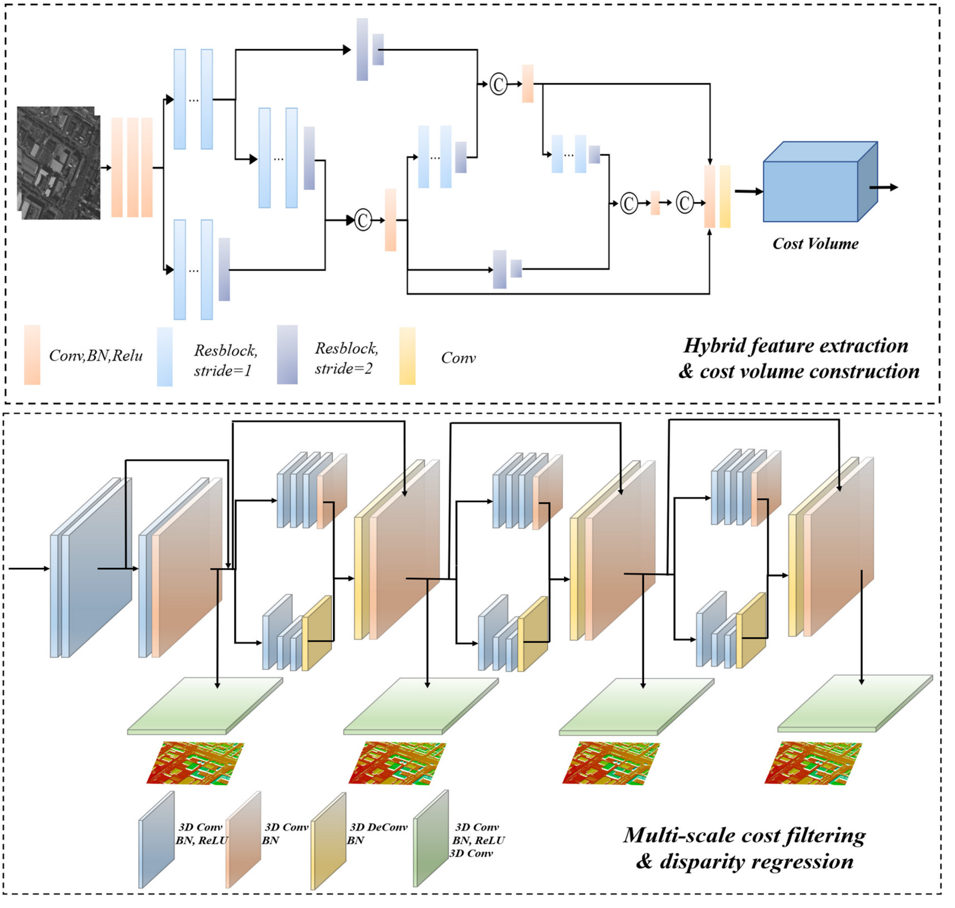


FIGURE 1 Detailed architecture of HF^{2}Net. The main pipeline includes hybrid feature extraction with cross-scale feature extractor (CSF), cost volume construction, multi-scale cost filtering and disparity regression.

Cross-scale feature extractor (CSF)

The ill-posed areas, such as texture-less, repetitive patterns, occluded buildings and disparity discontinuities regions, are always intractable for stereo matching methods. This problem is more severe on satellite stereos due to their broad coverage and complicated image content. Therefore, improving matching capability on intractable regions is effective for enhancing DLSMs' performance. To achieve this, we designed a CSF. Theoretically, local image features are sufficient for correct disparity estimation in texture-rich areas, while the intractable regions require more global context information (Huang et al., 2018; Xu et al., 2022). Thus, CSF adopts two asymmetric branches to differentiate structural-context features, which are named as structural branch and context branch. The structural branch captures local and prominent features to ensure robust matching in texture-rich areas. The context branch enlarges the repetitive field and aggregates global context information. Since long-range features are aggregated, the context branch can facilitate matching performance on intractable areas.

As shown in Figure 1a, we designed the hybrid feature extractor by stacking multiple CSF modules, whose detailed settings are listed in Table 1. The feature extraction module begins with three small convolution filters (3 × 3) to capture preliminary image features. The output feature map is labelled as $conv_0$. We then used a CSF module for higher level feature extraction, and the outputs of both the structural- and context branches are labelled as

TABLE 1 Settings of the hybrid feature extraction module of HF[2]Net.

Layers	Input	Parameters	Dimension
Input	–	–	$H \times W \times I_c$
$Conv_{0\_0}$	I_{ref} or I_{mat}	$\begin{bmatrix} 3\times3, I_c \rightarrow 32, s=1 \\ 3\times3, 32 \rightarrow 32, s=1 \\ 3\times3, 32 \rightarrow 32, s=1 \end{bmatrix}$	$H \times W \times 32$
$Conv_{0\_1}$	$Conv_{0\_0}$	$\begin{bmatrix} ResBlock, 32 \rightarrow 32, b=2, s=1 \\ ResBlock, 32 \rightarrow 32, b=1, s=2 \end{bmatrix}$	$\frac{1}{2}H \times \frac{1}{2}W \times 32$
$Conv_{0\_2}$	$Conv_{0\_0}$	$\begin{bmatrix} ResBlock, 32 \rightarrow 32, b=2, s=1 \\ ResBlock, 32 \rightarrow 32, b=1, s=2 \\ ResBlock, 32 \rightarrow 32, b=2, s=1 \\ ResBlock, 32 \rightarrow 64, b=1, s=2 \end{bmatrix}$	$\frac{1}{4}H \times \frac{1}{4}W \times 64$
$Conv_{1\_2}$	$Conv_{0\_1}$	$\begin{bmatrix} ResBlock, 32 \rightarrow 32, b=4, s=1 \\ ResBlock, 32 \rightarrow 64, b=1, s=2 \end{bmatrix}$	$\frac{1}{4}H \times \frac{1}{4}W \times 64$
$Conv_{1\_3}$	$Conv_{0\_1}$	$\begin{bmatrix} ResBlock, 32 \rightarrow 32, b=4, s=1 \\ ResBlock, 32 \rightarrow 32, b=1, s=2 \\ ResBlock, 32 \rightarrow 32, b=4, s=1 \\ ResBlock, 32 \rightarrow 64, b=1, s=2 \end{bmatrix}$	$\frac{1}{8}H \times \frac{1}{8}W \times 64$
$Conv_{2\_2}$	$Conv_{0\_2}$ $Conv_{1\_2}$	Concatenation and dimension transformation $1 \times 1128 \rightarrow 64, s = 1$	$\frac{1}{4}H \times \frac{1}{4}W \times 64$
$Conv_{2\_3}$	$Conv_{2\_2}$	$\begin{bmatrix} ResBlock, 64 \rightarrow 64, b=2, s=1 \\ ResBlock, 32 \rightarrow 128, b=1, s=2 \end{bmatrix}$	$\frac{1}{8}H \times \frac{1}{8}W \times 128$
$Conv_{2\_4}$	$Conv_{2\_2}$	$\begin{bmatrix} ResBlock, 64 \rightarrow 64, b=2, s=1 \\ ResBlock, 64 \rightarrow 64, b=1, s=2 \\ ResBlock, 64 \rightarrow 64, b=2, s=1 \\ ResBlock, 64 \rightarrow 128, b=1, s=2 \end{bmatrix}$	$\frac{1}{16}H \times \frac{1}{16}W \times 128$
$Conv_{3\_3}$	$Conv_{1\_3}$ $Conv_{2\_3}$	Concatenation and dimension transformation $1 \times 1192 \rightarrow 128, s = 1$	$\frac{1}{8}H \times \frac{1}{8}W \times 128$
$Conv_{3\_4}$	$Conv_{3\_3}$	$\begin{bmatrix} ResBlock, 128 \rightarrow 128, b=2, s=1 \\ ResBlock, 128 \rightarrow 192, b=1, s=2 \end{bmatrix}$	$\frac{1}{16}H \times \frac{1}{16}W \times 192$
$Conv_{4\_4}$	$Conv_{2\_4}$ $Conv_{3\_4}$	Concatenation and dimension transformation $1 \times 1320 \rightarrow 192, s = 1$	$\frac{1}{16}H \times \frac{1}{16}W \times 192$
$Conv_{4\_u}$	$Conv_{4\_4}$	Up-sampling: scale = 4	$\frac{1}{4}H \times \frac{1}{4}W \times 192$
$Conv_{3\_u}$	$Conv_{3\_3}$	Up-sampling: scale = 2	$\frac{1}{8}H \times \frac{1}{8}W \times 128$
$Conv_{4\_u}$	$Conv_{4\_4}$	Up-sampling: scale = 4	$\frac{1}{4}H \times \frac{1}{4}W \times 192$
$Conv_{final}$	$Conv_{0\_2}$ $Conv_{1\_2}$ $Conv_{3\_u}$ $Conv_{4\_u}$	Concatenation and dimension transformation $\begin{bmatrix} 3\times3{,}448 \rightarrow 128, s=1 \\ 1\times1{,}128 \rightarrow 32, s=1 \end{bmatrix}$	$\frac{1}{4}H \times \frac{1}{4}W \times 32$

Note: I_{ref} and I_{mat} refer to the reference image and matching image, respectively; and H and W refer to their height and width, respectively. I_c is the channel of the reference image, which is one or three in our experiments; b is the number of Resblock (He et al., 2016); s refers to stride; and sign $\rightarrow$ indicates the transform operation on the channel dimension.

*conv*0 _ 1 and *conv*0 _ 2. Another CSF is then applied on *conv*0 _ 1, whose outputs are *conv*1 _ 2 and *conv*1 _ 3. Note that *conv*0 _ 2 and *conv*1 _ 2 are equal in size, so we concatenated as *conv*2 _ 2. Since *conv*0 _ 2 and *conv*1 _ 2 come from different feature branches and are asymmetrical in local and global feature representation, their concatenation enables aggregation of multi-scale spatial feature cues and thus meets the goal of hybrid feature incorporation. Next, similar operations are applied to *conv*2 _ 2 to obtain *conv*2 _ 3 and *conv*2 _ 4. Then, *conv*3 _ 3 is naturally obtained by concatenating *conv*1 _ 3 and *conv*2 _ 3. Two ResBlocks (He et al., 2016) are then stacked to further capture global context information and down-sample *conv*3 _ 3 to *conv*3 _ 4. At last, the output *conv*3 _ 4 is concatenated with *conv*2 _ 4 for *conv*4 _ 4. In addition, we stack *conv*2 _ 2, up-sampled *conv*3 _ 3 and *conv*4 _ 4 to further facilitate different-level aggregation. In a word, the hybrid feature extractor sufficiently aggregates multi-scale features and hybrid structural-context information, thus facilitating matching correctness and robustness of DLSMs simultaneously.

The feature extraction module is applied to the left and right images with shared weights for unary feature extraction. The extracted hybrid features of the stereo images are then concatenated following the rules introduced by PSMNet (Chang & Chen, 2018), resulting in a 4D cost volume with the size of $C \times \frac{1}{4}D_{\text{max}} \times \frac{1}{4}H \times \frac{1}{4}W$, where C, D_{max}, H and W refer to the number of feature maps, preset disparity range, image height and width, respectively. The cost volume counts the possibility of different disparity values in each pixel and then feeds them into the cost filtering module for disparity optimization.

Multi-scale cost filtering

With the 4D cost volume as input, we first use four small 3D convolution filters ($3 \times 3 \times 3$) to rectify the cost values. This operation filters out most mismatching values and determines the optimal ones of each pixel. Subsequently, the initial disparity map is generated via bilinear interpolation and disparity regression.

To deal with ill-posed areas and further regularize the cost volume, PSMNet directly stacks the 3D convolutions and the transposed 3D convolutions, then skip-connects three cost volumes of the same resolution to build a stack-hourglass structure for cost filtering. This kind of structure exploits more global context information while missing detailed information. Unlike PSMNet, the proposed cost filtering module preserves more details by aggregating multi-scale features. As shown in Figure 1b, asymmetric branches are taken to build the cost filtering module, where one aggregates local cost and preserves image details, and the other exploits more context information. The detailed settings of the proposed multi-scale cost filtering module are listed in Table 2. The low- and high-level branches down-sample the cost volume by two and four times, respectively. Both branches are up-sampled and concatenated with the preliminary regularized cost volume for further cost optimization. Considering the matching accuracy and efficiency, we adopt three cost filtering modules for cost optimization, obtaining three regularized cost volumes.

Disparity estimation

The size of all the regularized cost volumes is $\frac{1}{4}D_{\text{max}} \times \frac{1}{4}H \times \frac{1}{4}W$. With the given cost volumes, we estimated low-resolution disparity maps and then up-sampled them to image size. Since the regression-based method is more robust than the classification-based methods and can retain sub-pixel accuracy, we use a soft-argmin operation $\varphi(\bullet)$, as described in Equation (1), to estimate disparity maps with continuous disparity values. The probability of each disparity level d is predicted and the estimated disparity value $\hat{d}$ is counted as the sum of all the disparity levels weighted by their probabilities p:

$$\hat{d} = \sum_{0}^{D_{\text{max}}} d \times \varphi(p). \tag{1}$$

TABLE 2 Settings of the multi-scale cost filtering module of HF^{2}Net.

Layers	Input	Parameters	Dimension
Input	–	–	$C \times D_{\text{range}} \times \frac{1}{4}H \times \frac{1}{4}W$
Primarily cost filtering			
$3DConv_0$	Initial cost volume	$\begin{bmatrix} 3\times3, 64 \rightarrow 32, s=1 \\ 3\times3, 32 \rightarrow 32, s=1 \end{bmatrix}$	$32 \times D_{\text{range}} \times \frac{1}{4}H \times \frac{1}{4}W$
$3DConv_1$	$3DConv_0$	$\begin{bmatrix} 3\times3, 32 \rightarrow 32, s=1 \\ 3\times3, 32 \rightarrow 32, s=1 \end{bmatrix}$	$32 \times D_{\text{range}} \times \frac{1}{4}H \times \frac{1}{4}W$
Multi-scale cost filtering ($x = 2, 3, 4$ in $3DConv_x$, respectively)			
$3DConv_{x\_l}$	$3DConv_{x-1}$	$\begin{bmatrix} 3\times3, 32 \rightarrow 32, s=2 \\ 3\times3, 32 \rightarrow 64, s=1 \\ 3\times3, 64 \rightarrow 64, s=1 \end{bmatrix}$	$64 \times D_{range} \times \frac{1}{8}H \times \frac{1}{8}W$
$3DConv_{x\_h}$	$3DConv_{x-1}$	$\begin{bmatrix} 3\times3, 32 \rightarrow 64, s=2 \\ 3\times3, 64 \rightarrow 64, s=2 \\ 3\times3, 64 \rightarrow 64, s=1 \\ 3\times3, 64 \rightarrow 64, de\_s=2 \end{bmatrix}$	$64 \times D_{range} \times \frac{1}{8}H \times \frac{1}{8}W$
$3DConv_{x\_l}$	$3DConv_{x\_l}$	$1 \times 1, 64 \rightarrow 64, s = 1$	$64 \times D_{\text{range}} \times \frac{1}{8}H \times \frac{1}{8}W$
Agg_1	$3DConv_{x\_l}$ $3DConv_{x\_h}$	Element-level summation	$64 \times D_{\text{range}} \times \frac{1}{8}H \times \frac{1}{8}W$
Agg_1	*Agg_1*	$1 \times 1, 64 \rightarrow 32, de\_s = 2$	$32 \times D_{\text{range}} \times \frac{1}{4}H \times \frac{1}{4}W$
$3DConv_{x\_re1}$	Agg_1 $3DConv_{x-1}$	Element-level summation	$32 \times D_{\text{range}} \times \frac{1}{4}H \times \frac{1}{4}W$

Note: D_{range} refers to the preset disparity range.

Four disparity maps $(\text{Dis}_0, \text{Dis}_1, \text{Dis}_2, \text{Dis}_3)$ with a size of $H \times W$ are obtained in this step. During the training phase, the difference between them and the given ground truth are parallelly computed to formula the final loss for model optimization. During the testing phase, disparity map Dis_3 is selected as the final output.

Loss function

Due to its robustness and low sensitivity to outliers, we adopt the widely used smooth L_1 loss function to supervise the training process of HF^{2}Net. The loss function is defined as Equation (2):

$$\mathcal{L}\left(\hat{d}\right) = \sum_{n=1}^{N} \text{smooth}_{\ell 1}\left(d_n^{pre}, d_n^{gt}\right) \tag{2}$$

in which:

$$\text{smooth}_{\ell 1}(d) = \begin{cases} 0.5d^2, \text{if } |d| < 1 \\ |d| - 0.5, \text{otherwise} \end{cases} \tag{3}$$

where N is the number of labelled pixels, d_n^{gt} is the ground truth disparity and d_n^{pre} is the estimated disparity. The total loss of HF^{2}Net is a weighted summation of different stages, which is defined as Equation (4):

$$\mathcal{L} = \sum_{i=0}^{i=3} \lambda_i \mathcal{L}_i \tag{4}$$

where λ_0, λ_1, λ_2, and λ_3 are the weights of Dis_0, Dis_1, Dis_2, and Dis_3, respectively.

DLSMs-based workflow for large-scale DSM generation

Overall workflow

Figure 2 displays the complete procedure of the proposed DSM generation workflow, including pre-processing, disparity estimation and post-processing. The processing step takes satellite stereos with corrected rational polynomial coefficients (RPC) as the input and shifts unmatched disparity ranges into a fixed scope and then crops satellite stereos to regular patches. The disparity estimation step comprises training phase and testing phase. In the training phase, the proposed HF^{2}Net is trained with high-accuracy SISM datasets for parameters optimization. The well-trained HF^{2}Net then predicts disparity maps for each satellite patch. At last, the post-processing step generates initial DSM patches, then refines and merges them to obtain large-scale DSM.

Pre-processing

The pre-processing serves for epipolar correction, disparity shifting and patch crop. The DLSMs have a lower tolerance to the y-parallax of satellite stereos than conventional methods since the cost volume only takes pixels at the same row into consideration. Therefore, we use the SRTM-aided epipolar resampling method (Hu et al., 2019) as the default setting for epipolar correction because it can limit the y-parallax of satellite stereos to the 0.5 pixels level.

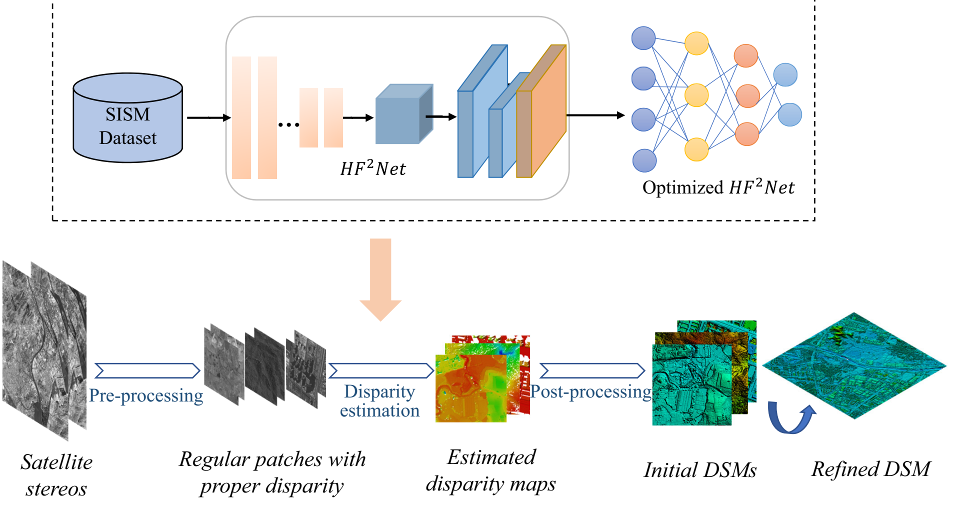


FIGURE 2 The general procedure of the DSM generation workflow.

Satellite stereos with broad coverage always have a wide disparity scope, especially when numerous steep mountains and tall buildings exist. However, most DLSMs can only process a relatively narrow disparity range due to the use of 3D convolutions. The unmatched disparity range results in severe matching errors when applying DLSMs to satellite stereos and needs to be pre-processed. To do this, we shift the disparity of satellite stereos with the assistance of a low-resolution DSM or DEM (i.e., SRTM) to meet the requirements of DLSMs and then crop the epipolar stereos to regular patches. The processing steps are as follows:

- We first segment epipolar stereos into several independent parts according to their terrain slope. Thus, the disparity ranges are roughly determined within the same terrain rather than the whole satellite stereos. This step narrows down the disparity range for some geographical scenes such as the flat regions.
- We then calculate each part's minimum and maximum disparity D_{min} and D_{max} with the given low-resolution DSM/DEM. Since the low-resolution DSM/DEM can only provide a rough disparity range and may be inaccurate if changes occur, we give offsets ΔD_{min} and ΔD_{max} to the D_{min} and D_{max}. That is, we take $\left[D_{min} - \Delta D_{min},\ D_{max} + \Delta D_{max}\right]$ as the initial disparity range for disparity shifting. Here, the offsets are empirically determined by the base-height ratio of satellite stereos and the elevation change ΔH.
- Next, we shift the minimum disparity $D_{min} - \Delta D_{min}$ to 0 at first, then all the disparities in the processing part are adjusted according to their distance to $D_{min} - \Delta D_{min}$. At last, we crop the reference images to regular patches and then crop the matching images with the shifted disparities.

Disparity estimation

The disparity estimation phase takes cropped patches into the trained HF^{2}Net and estimates their disparity values. It should be noted that we use the shifted disparity range for all patches from the same part, that is $\left[0,\ \left(D_{max} + \Delta D_{max}\right) - \left(D_{min} - \Delta D_{min}\right)\right]$, rather than determine specific a disparity range for each patch during the disparity estimation process.

Post-processing

The post-processing phase first generates initial DSM patches with the estimated disparity, then refines the generated DSM. The HF^{2}Net gives estimated disparity values for all pixels of the reference patches. Thus, the left–right consistency check should be performed on the estimated disparity maps to filter out occlusion pixels and some matching errors.

Consistency check

Theoretically, the pixel p_{ref} in the reference image has a unique correspondence p_{mat} in the matching image. Therefore, the conventional methods take twice the matches by using the reference image and the matching image as references, respectively, and then compare the distance difference of $|\ p_{ref} - p_{mat}\ |$ to formulate the left–right consistency check. However, this does not always work for DLSMs since the disparity estimated by networks does not strictly follow the left–right consistency rule. Therefore, we use the initial DSMs generated with the network estimated disparities as reference to accomplish the process of left–right consistency check as well as occlusion detection. We follow the basic idea that the generated points by the pixels p_{ref} and p_{mat} can be back-projected for visible and correct matching pixels while not for the occluded pixels. Thus, the specific processing steps are as follows:

- Generate the initial DSMs with the estimated disparities. Since the occlusion pixels and matching errors are not filtered out, two kinds of points exist in the generated DSMs. That is, the correct points come from visible and correct matching pixels and the wrong ones come from occluded or mismatching pixels. Thus, the objective is to eliminate the wrong points from the initial DSMs.
- Project the points of the initial DSMs to the reference images and the matching ones. Labelled the projected points as p_{ref} and p_{check}, then calculated new disparity maps with p_{ref} and p_{check}.
- As a basic rule, the points that come from p_{ref} and p_{mat} can be projected back for the visible and correct matching pixels, while cannot reach p_{mat} for the occluded or mismatching ones. That is, $|q_{\mathrm{check}} - q_{\mathrm{mat}}| < T$ should be valid for the visible and correct matching pixels but not for the occluded or mismatched ones. Given a small threshold T, we filter out all p_{mat} where $|q_{\mathrm{check}} - q_{\mathrm{mat}}| < T$ is not valid.

DSM refinement

The above-mentioned process removes occluded and mismatched pixels from the estimated disparity maps, resulting in some holes in the re-generated DSMs. We complete the holes with bilinear interpolation. Besides, we further refine water regions under the supervision of the global surface water (Pekel et al., 2016). At last, we apply bilateral filter and median filter on the DSMs, where the former removes salt and pepper noise and the latter eliminates outliers.

EXPERIMENTAL ANALYSIS

In this section, we make a detailed analysis of the proposed HF^{2}Net and the DSM generation workflow. Since the performance of HF^{2}Net determines the results of the proposed DSM generation workflow, we first investigated its optimal structure and default settings. We then evaluated HF^{2}Net's accuracy on benchmark datasets, from disparity estimation to primary DSM generation. Next, we implemented the whole workflow on two representative GF-7 stereos to present its robustness and practical capability. At last, we comprehensively discussed the superiority and shortcomings of the proposed DSM generation workflow.

Optimal settings determination

The experiments to determine optimal network settings include ablation studies and weight settings. The ablation studies illustrated the effectiveness of the designed model components, and the weight-setting experiments investigated the optimal weight settings for different outputs.

Implementation details

We conducted ablation studies and optimal-setting experiments on the US3D dataset (Bosch et al., 2016). The US3D dataset comprises 4292 RGB samples collected by the WorldView-3 satellite, with a resolution of 0.3 m and size of 1024×1024 pixels. We randomly selected 80% samples for training and the rest 20% for validation. All models were implemented on Pytorch platform using four NVIDIA Tesla V100 GPUs. We used Adam as the optimizer, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. During the training phase, images were randomly cropped to 512×512 pixels. The disparity range was −128 to 128, and the batch size was set as 4. We trained each model from scratch for 50 epochs. The initial learning rate was 0.004 and dropped to half at epoch 16 and 30.

Ablation study

We took PSMNet (Chang & Chen, 2018) as the baseline model. By replacing part of PSMNet's components, we evaluated the effectiveness of the designed modules, including the hybrid feature extraction module, the cost filtering module and its numbers. As listed in Table 3, We marked the feature extraction module of PSMNet and HF^2Net as *PSM_Fea* and *HF^2_Fea*, respectively. Their cost filtering modules were labelled as *PSM_3D* and *HF^2_i_3D* ($i = 0, 1, 2, 3$), where i represented the numbers of the designed multi-scale cost filtering module. We divided the experiments into three groups for better illustration. The evaluation metrics are the average end-point error (EPE) of all validation samples, which are expressed as follows:

$$\mathrm{EPE} = \frac{1}{N} \sum_{i \in N} | d_i^{\mathrm{pre}} - d_i^{gt} | \tag{5}$$

where N represents sample numbers of the validation set, d_i^{pre} and d_i^{gt}represent the predicted disparity map and the given label of the i-th sample, respectively.

Experiments in Group 1 illustrated the effectiveness of the proposed hybrid feature extraction module. The EPE of PSMNet was 1.579 on the US3D dataset. By replacing *PSM_Fea* with *HF^2_Fea*, the EPE significantly decreased to 1.440. Experiments in Group 2 showed the significance of the proposed cost filtering module. It could be noticed that the network did not converge when no multi-scale cost filtering modules were applied, no matter using *PSM_Fea* or *HF^2_Fea*. According to experiments in Groups 1 and 2, we concluded that the hybrid feature extraction and multi-scale cost filtering modules are effective for the proposed HF^2Net. According to experiments in Groups 1 and 2, we concluded that both the hybrid feature extraction module and the multi-scale cost filtering module are effective for the proposed HF^2Net.

Optimal-setting experiments

The ablation study indicated that the optimal network structure would be obtained with three multi-scale cost filtering modules. Thus, we further investigated the influence of different weight settings among various weight combinations. Suppose the disparity maps from the preliminary cost volume and the three multi-scale cost filtering modules as Dis_0, Dis_1, Dis_2, and Dis_3, we marked their weights as λ_0, λ_1,λ_2, and λ_3, respectively. Although the

TABLE 3 Ablation study of HF^2Net.

Feature extractor		Cost filter					
PSM_Fea	HF^2_Fea	PSM_3D	HF^2_0_3D	HF^2_1_3D	HF^2_2_3D	HF^2_3_3D	EPE
Group 1							
√		√					1.579
	√	√					1.440
Group 2							
√			√				3.318
	√		√				3.076
Group 3							
	√			√			1.433
	√				√		1.462
	√					√	**1.390**

Note: The bold value indicates the optimal accuracy obtained with the settings.

preliminary cost volume is effective for disparity optimization, it is less reliable than the final regularized cost volume. Therefore, we empirically assumed that it had a relatively smaller weight. Since Dis_3 is the ultimate output in the testing phase, we gave it a larger weight. Table 4 displays the experimental results with different weight settings. It shows that the lowest EPE was yielded with the setting $\lambda_0 = 0.5$, $\lambda_1 = 0.5$, $\lambda_2 = 0.7$, $\lambda_3 = 0.7$, which was 1.390 on the US3D dataset. It should be noted that we observe no significance performance difference when applying different weight settings.

Performance evaluation of HF^{2}Net

In this section, we first conducted experiments on the US3D dataset to observe HF^{2}Net's performance on disparity estimation by comparing it with several state-of-the-art methods. Since the US3D dataset does not provide images' RPC information, we then conducted the DSM generation experiments with the grss_dfc_2019 data (http://www.grss-ieee.org/community/technical-committees/data-fusion) to observe the performance of the generated DSMs further (Bosch et al., 2019; Le Saux et al., 2019). We noted that the provided samples of grss_dfc_2019 data were cropped from the same satellite stereos as the US3D dataset.

Implementation details

The US3D dataset was randomly divided into three parts, with 3600 samples for training, 400 for validation and the remaining 292 for testing. To evaluate the performance of HF^{2}Net, we selected three representative methods for comparison: SGM (Hirschmuller, 2007), PSMNet (Chang & Chen, 2018) and ACVNet (Xu et al., 2022). SGM is a typical and widely used conventional stereo matching algorithm. PSMNet is widely studied and has always served as the baseline for DLSMs. ACVNet is one of the top-ranking DLSMs on the KITTI benchmark. We trained PSMNet, ACVNet, and the proposed HF^{2}Net under the same settings to avoid the influence of different training strategies. All the parameters were set as introduced in Section 4.1.1. When all the networks converged, we selected their optimal models for disparity prediction. We used EPE and the fraction of erroneous pixels ($D1$) for accuracy indicators, where $D1$ is expressed as follows:

$$D1 = \frac{1}{N} \sum_{n \in N} \left[| d_n^{pre} - d_n^{gt} | > T \right] \tag{6}$$

where N represents sample numbers of the validation set, d_n^{pre} and d_n^{gt} represent the predicted disparity map and the given label of the n-th sample, and T represent the threshold of error parallax and was set as 3 in our experiments.

TABLE 4 . Evaluation of different weight settings.

λ_0	λ_1	λ_2	λ_3	EPE
0.5	0.5	0.5	0.5	1.512
0.5	0.5	0.7	0.7	**1.390**
0.5	0.5	0.7	1.0	1.458
0.5	0.7	0.7	0.7	1.429
0.5	0.7	0.7	1.0	1.404

Note: The bold value indicates the optimal accuracy obtained with the settings.

TABLE 5 Accuracy comparison on the US3D dataset.

	SGM		PSMNet		ACVNet		HF²Net	
	EPE	D1	EPE (m)	D1	EPE	D1	EPE	D1
Testing set	12.78	22.55	1.53	9.66	1.52	9.95	**1.46**	**9.13**

Note: Optimal results are shown in bold.

TABLE 6 Accuracy comparison of three representative samples on the US3D dataset.

	SGM		PSMNet		ACVNet		HF²Net	
	EPE	D1	EPE	D1	EPE	D1	EPE	D1
No. OMA144	26.45	83.3	3.25	65.9	3.70	86.6	**3.13**	**62.5**
No. JAX 416	9.62	18.97	2.93	16.6	**2.80**	**15.98**	**2.80**	16.21
No. OMA287	11.70	12.42	0.84	4.07	0.90	3.88	**0.81**	**3.57**

Note: Optimal results are shown in bold.

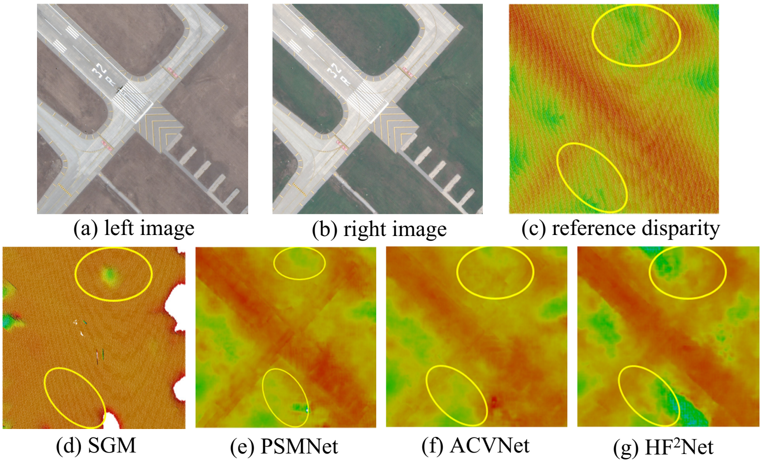


FIGURE 3 Visualization of the disparity maps generated by different methods (No. OMA 144).

Disparity estimation on the US3D dataset

Table 5 shows the overall accuracy of the 292 testing samples. It can be noticed that all the DLSMs significantly outperform SGM. For example, the EPE and *D*1 of SGM are 12.78 pixels and 22.55%, while the values of HF^{2}Net are only 1.46 pixels and 9.13%. The experimental results indicate that DLSMs can dramatically alleviate the mismatching problem of conventional methods and achieve much better results. Among the three DLSMs, HF^{2}Net yields the best performance, demonstrating its superiority.

We selected three representative samples from the testing set to illustrate the performance difference: OMA 144, JAX 416 and OMA 287. Among the three samples, OMA 144 represents intractable regions, JAX 416 and OMA 287 represent building scenes. Table 6 shows the quantitative results of the selected samples, and Figures 4 and 5 display the predicted disparity maps.

The accuracy results in Table 6 are in line with the overall accuracy shown in Table 5, that is, the DLSMs outperform SGM. ACVNet performs best on sample JAX416, with 2.8 pixels EPE and 15.98% D1. HF^{2}Net has equal EPE to ACVNet, while its D1 is 0.23% larger than ACVNet. The proposed HF^{2}Net gains the best performance on

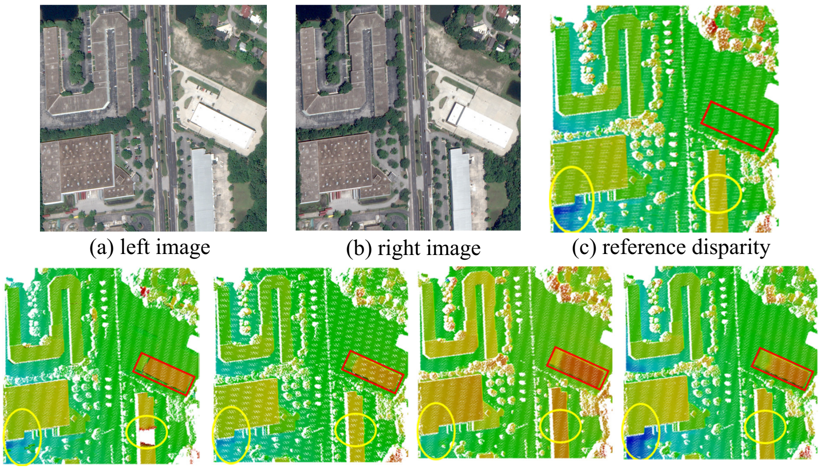


FIGURE 4 Visualization of the disparity maps generated by different methods (No. JAX 416).

FIGURE 5 Visualization of the disparity maps generated by different methods (No. OMA 287).

OMA 144 and OMA 287, with EPEs of 3.13 and 0.81 pixels on the two samples. It should be noted that though the accuracy of DLSMs significantly surpasses SGM, they face severe mismatching problems on the sample OMA 144 (Table 6 and Figure 3). According to Table 6, even the best-performed HF^{2}Net still has 62.5% D1, indicating that the majority of the scene was mismatched. Figure 3 shows the visualization results of the scene. It can be found that though severe mismatch occurs, the approximate shape of runways can be distinguished from the DLSMs' results, which cannot be seen from the results of SGM.

Table 6 also shows an accuracy gap between JAX 416 and OMA 287 though both are building scenes. Since the EPE and D1 on the sample JAX 416 seem anomalous, we carefully analyse the estimated disparity maps in Figure 4. As shown in the red box of Figure 4, there is a building in the images and all the methods predict its disparity correctly. However, the reference disparity map does not include this building. Thus, the inaccurate reference explains the anomalous results, providing the insight that the actual accuracy of JAX 416 should be better than those reported in Table 6. Disregarding this area, most regions were correctly predicted with similar disparity values except for the two ellipsoid regions. It can be observed that the result of HF^{2}Net in the left ellipsoid is the closest to the reference, indicating its better performance over the other methods. In addition, only SGM mismatched part of the building in the right ellipsoid, explaining why its accuracy is lower than the other methods.

Figure 5 shows the matching results in dense building scenes, where severe occlusion and disparity discontinuities always occur. As shown in the marked regions, the SGM misses disparity estimation at the occlusion and disparity discontinuities regions. Thus, there are no disparity values in the disparity map. Unlike SGM, the DLSMs still estimate disparity values at the occluded and discontinuity regions.

DSM generation

Table 7 and Figure 6 illustrate the quantitative and qualitative assessment of the generated DSMs. In Table 7, the root mean square error (RMSE) and mean error (ME) between the generated DSMs and the given reference were computed, as well as the average results of the whole testing set.

According to Table 7, HF^{2}Net gains the highest accuracy on the whole testing set and most of the selected samples. The RMSE and ME indicators of HF^{2}Net are only 3.77m and 2.14m, while the values of the second-best

TABLE 7 Accuracy comparison of the generated DSMs on the grss_dfc_2019 data.

	SGM		PSMNet		ACVNet		HF^{2}Net	
	RMSE	ME	RMSE	ME	RMSE	ME	RMSE	ME
JAX_068_DSM	4.75	2.03	5.04	2.04	4.64	2.06	**3.56**	**1.27**
JAX_251_DSM	3.42	1.97	3.24	1.82	**3.10**	**1.69**	3.42	2.03
JAX_467_DSM	3.20	1.57	2.50	1.26	2.5	1.37	**2.41**	**1.10**
Testing set	9.90	4.73	5.55	3.64	5.59	3.80	**3.77**	**2.41**

Note: Optimal results are shown in bold.

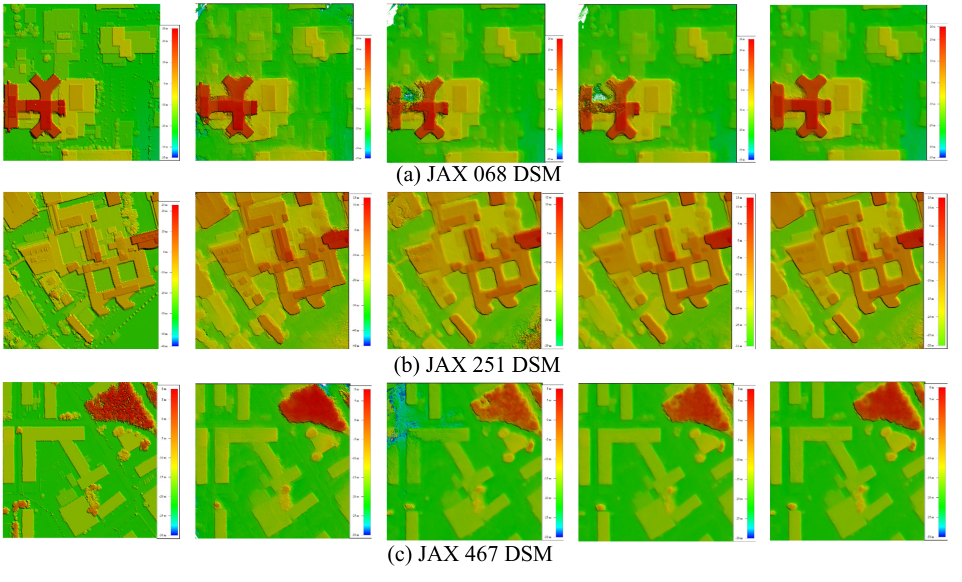


FIGURE 6 The generated DSMs of three samples from the grss_dfc_2019 data (in each line, from left to right, are the reference DSMs, the generated DSMs of SGM, PSMNet, ACVNet, and HF^{2}Net, respectively).

PSMNet are 5.55m and 3.64 m, demonstrating the superiority of the proposed HF^{2}Net. In most cases, SGM still performs worse than DLSMs, while the accuracy differences were not as significant as the difference on disparity maps.

Figure 6 shows the generated DSMs of the selected three samples. Most of the generated DSMs are correctly reconstructed, which can be attributed to the size of the given block DSMs being relatively small, and the selected areas are those with distinctive features, such as low-rise buildings. Compared with the DLSMs, the DSMs generated by SGM have more feature details (e.g., small vegetation around buildings). Among the three DLSMs, there is a slight difference between the DSMs generated by ACVNet and HF2Net, while the results of PSMNet are worse than the other two DLSMs.

Implementation for large-scale DSM generation

The experiments in Sections 4.1 and 4.2 investigate the optimal network structure of HF^{2}Net and primarily demonstrate its effectiveness for disparity estimation and DSM generation. In this section, we embedded the optimized HF^{2}Net model into the proposed DSM generation workflow for disparity estimation to further evaluate its practical capability for large-scale DSM generation.

Implementation details

Training dataset

Since the US3D dataset and grss_dfc_2019 data do not provide the large-scale satellite stereos, we conducted the complete DSM generation experiments on GF-7 satellite stereos. GF-7 is Chinese first civilian stereo mapping satellite that was launched on 3 November 2019. It is equipped with a two-line camera that can catch stereo panchromatic images at the same time (Xie et al., 2020). The resolution of GF-7 is 0.65 m for the backward images and 0.8 m for the forward images, respectively. To better exploit the workflow's potential and avoid the domain adaptation problem of DLSMs, we prepared a training dataset with GF-7 stereos in advance to train DLSMs. The self-made dataset was created from four GF-7 pairs in the Guangdong province of China, with various complex scenes in a coverage range exceeding 2000 km^2. The images resolution is resampled to 0.8 m. The ground truth parallax is calculated with highly accurate LiDAR-derived DSM, whose resolution is 0.5 m. The dataset comprises 5400 panchromatic training samples with the size of 768 × 768 pixels. The disparity range of the whole training set is within 0–224 pixels.

Comparison methods

To measure the performance of HF^{2}Net, we still use PSMNet and ACVNet as DLSM comparisons in this section. For a fair comparison, we replaced SGM with SRDM (Huang et al., 2018). SRDM is one of the most advanced SGM modifications, which uses initial matching points as constraints to improve the matching accuracy in building areas and has been proven to be effective (Zhang, Zou, et al., 2022). It needs to mention that all the DLSMs were trained from scratch with the same parameters, and their optimal models were selected for DSM generation. The training process was the same as described in Section 4.1.1.

DSM generation settings

The DSM generation process was performed on a workstation with 64G RAM and an NVIDIA GeForce RTX3090 GPU with 24G GPU memory. In our experiment, the pre-processing step split satellite stereos into 16 regions and used SRTM for initial disparity range determination. Satellite stereos were then cut into regular patches with the size of 1536 × 2048 pixels. In the disparity estimation step, SRDM directly took the regular patches into its default settings for disparity estimation. All the trained DLSMs were first converted to executors by the LibTorch library and then used for disparity estimation. In the post-processing step, all the estimated disparity maps were forward intersected under the same RPC parameters for a fair comparison. All the refinement operations as described in Section 3.2 were then applied to obtain the final DSMs.

Accuracy assessment

To quantitively assess the generated DSM's performance, we used the ground-control points (GCPs) and LiDAR data-derived DSMs for accuracy statistics. For GCPs, we compared them with the points at the same location obtained from generated DSMs. With LiDAR data-derived DSM as a reference, we further computed each DSM's RMSE and ME against the reference DSM. Furthermore, we determined several residual intervals and counted the percentage of residuals in different intervals to further observe the residual distribution of the generated DSMs. Considering the elevation accuracy of different terrain types are different, we empirically determined several residual thresholds to illustrate the matching completeness better. With the given thresholds, we neglected the changed regions and false matching points from the accuracy statistics and thus could better depict the valid elevation residuals of these methods. In our experiments, the residual intervals were set as 0–1, 1–5, 5–10 and >10 m. The residual thresholds for large flat regions (texture-less areas and repetitive regions), building regions (occlusions and discontinuous disparity areas), and mountain regions were 1, 5 and 10 m, respectively.

Test data

The DSM generation experiment was conducted on another two GF-7 stereos, which locates in different positions with the prepared training set. Both GF-7 stereos were captured in 2021. The first stereo was captured in Zhongshan city, Guangdong province, with a coverage of 465 km^2. The second stereo were captured in Guangzhou city, Guangdong, with a coverage of 601 km^2. The topography of the two GF-7 stereos featured various terrain types and thus were suitable for comprehensive accuracy analysis. The first GF-7 stereo mainly featured flat regions with large tracts of farmlands and concentrated town residential areas. Two main rivers passed through the area and converged on the right side, with several low mountains distributed around the rivers. The second GF-7 stereo contained dense urban regions comprised of concentrated urban village areas and large and tall buildings, hilly areas, and large mountainous areas. For example, the high mountains in the left corner of the pair had an approximately 400 m relief difference. In our experiments, we selected six representative areas of interest (AOIs) from the two GF-7 pairs to assess the pipeline's performance comprehensively, including texture-less areas, concentrated residentials, and mountains. The details of the two stereos and the selected AOIs are displayed in Figures 7 and 8.

Accuracy assessment with GCPs

In each GF-7 stereo, we selected nine GCPs that were evenly distributed in rigid locations for accuracy assessment. Since satellite stereos were registered to precise locations with control data in advance, we directly compared the GCPs and the points extracted from the generated DSMs according to the GCPs' plane coordinates for accuracy assessment. The computed residuals are presented in Tables 8 and 9, respectively.

Table 8 presents the residuals of the first GF-7 stereo, where large texture-less areas and repetitive regions existed, such as large tracts of farmlands and concentrated town residential areas. All the methods performed well on these GCPs, while the elevation residuals of the DLSMs were less than SRDM on most GCPs, even with the baseline PSMNet. The experimental results were in line with previous works (Chang & Chen, 2018; He et al., 2022), where the DLSMs' performance was superior to the conventional methods when dealing with intractable regions. Specifically, the proposed HF^{2}Net produced residual values that were less than 1 m on eight GCPs and achieved the lowest residuals on five GCPs. These results demonstrated the effectiveness of the proposed CSF extractor

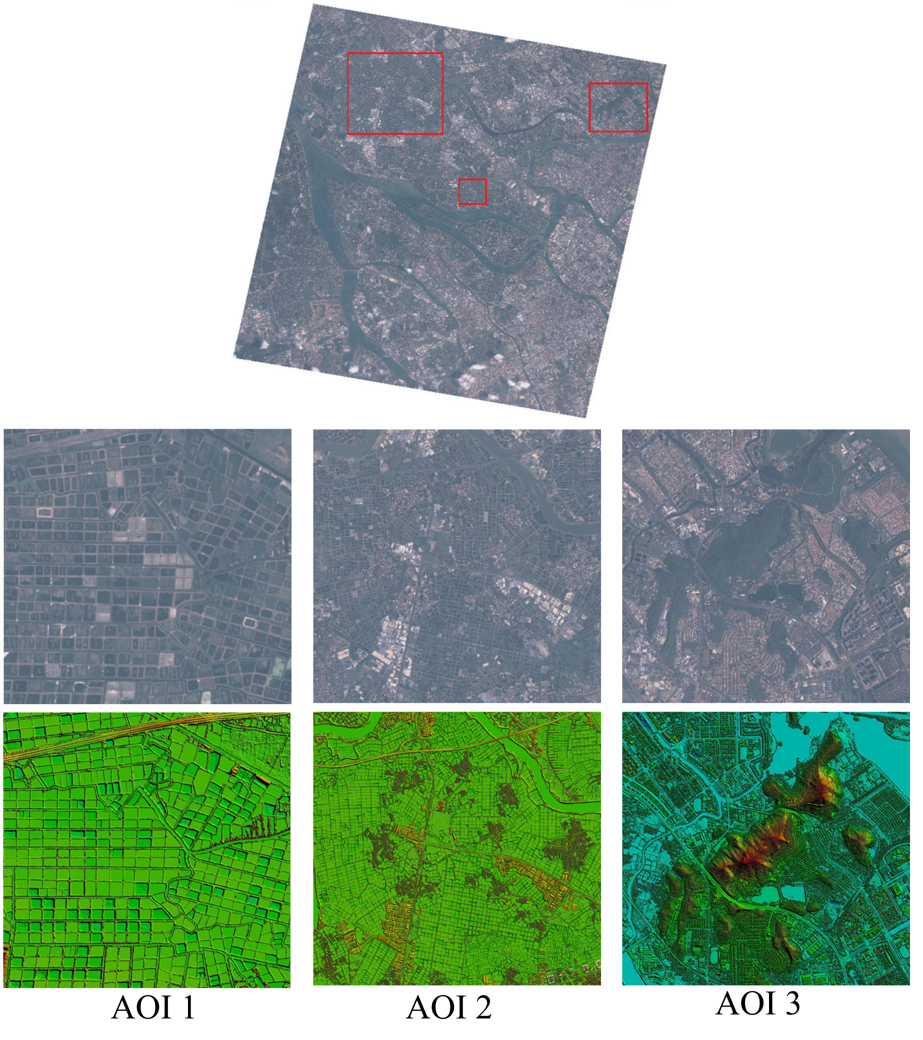


FIGURE 7 Visualization of the first GF-7 stereo, the orthophoto images and the reference LiDAR data-derived DSMs for AOI 1–AOI 3.

for intractable regions. Table 9 displays the elevation residuals on the second GF-7 stereo, which features a more complicated geographical scene than the first one. As shown in Table 9, HF^{2}Net and ACVNet outperformed SRDM on all GCPs while PSMNet did not show obvious superiority over SRDM. In addition, the proposed HF^{2}Net had the lowest elevation residuals on six GCPs. Tables 8 and 9 also indicate that HF^{2}Net and ACVNet have similar performance. Their elevation differences were less than 0.3 m in most cases, even at the centimetre level. Compared with the conventional method SRDM, the proposed HF^{2}Net consistently had lower residuals except for the seventh GCP in the first GF-7 stereo.

Accuracy assessment with the reference LiDAR data-derived DSMs

Six AOIs that featured different terrain types were selected to quantitively evaluate the overall performance of the proposed DSM generation workflow. AOI 1 and AOI 2 represented large flat regions, featuring texture-less areas and repetitive regions. AOI 4 and AOI 5 displayed building regions, where occlusion and disparity discontinuous

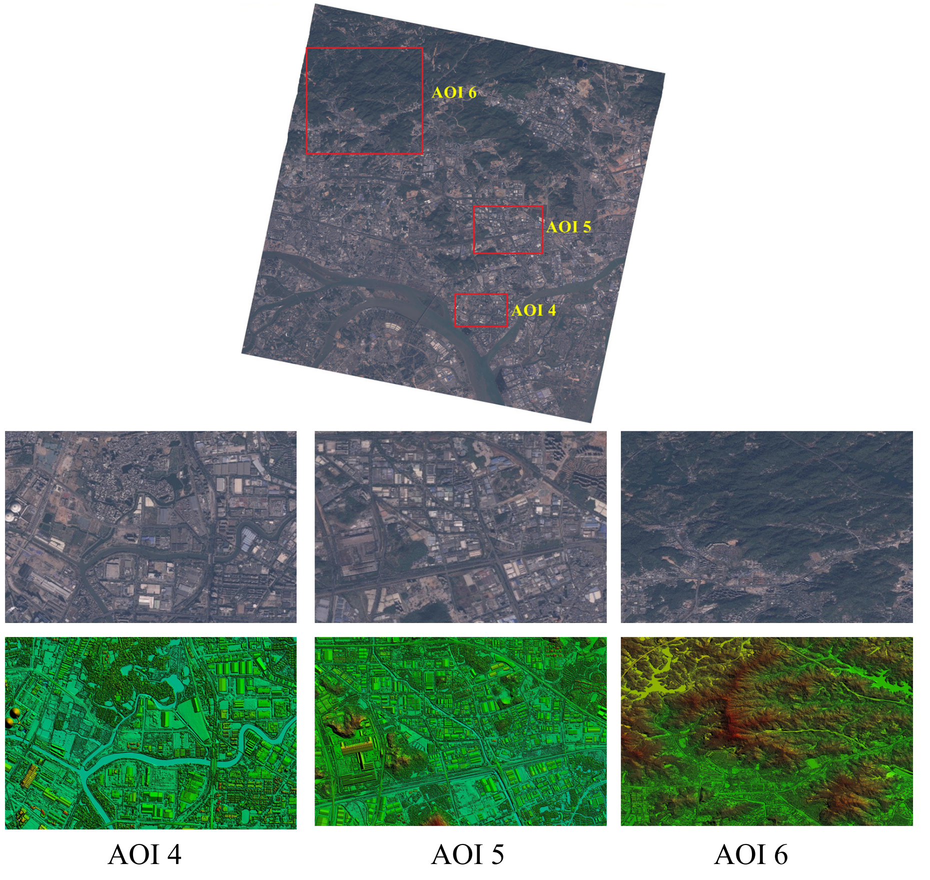


FIGURE 8 Visualization of the second GF-7 stereo, the orthophoto images and the reference LiDAR data-derived DSMs for AOI 4–AOI 6.

TABLE 8 Elevation residuals between the GCPs and the generated DSMs in the first GF-7 stereo (unit: meter).

	SRDM	PSMNet	ACVNet	HF^2Net
GCP 1	−1.76	*−1.92*	−1.06	−1.03
GCP 2	*−1.91*	−0.94	−0.80	−0.60
GCP 3	*−1.29*	−0.56	−0.13	−0.19
GCP 4	*−1.24*	−0.56	−1.14	−0.84
GCP 5	*0.19*	−0.10	0.05	−0.01
GCP 6	*−0.97*	−0.89	0.25	−0.10
GCP 7	−0.29	*−1.15*	−0.86	−0.64
GCP 8	*−2.72*	−0.91	−0.37	−0.54
GCP 9	*−1.04*	0.12	0.41	0.04

Note: Res = $Ele_{GCPs} - Ele_{gen}$, where Ele_{GCPs} and Ele_{gen} refer to the elevation of GCPs and the generated DSMs. The best results are underlined, while the worst are italicized.

cases always occurred. AOI 3 and AOI 6 featured mountains. In order to better illustrate the performance of the proposed workflow, we grouped the experimental results according to the terrain types and showed their quantitative results from Tables 10–12.

TABLE 9 Elevation residuals between the GCPs and the generated DSMs in the second GF-7 stereo (unit: meter).

	SRDM	PSMNet	ACVNet	HF^2Net
GCP 1	*−1.02*	0.89	0.20	0.20
GCP 2	5.25	*6.57*	3.77	3.35
GCP 3	−1.51	*3.38*	−0.38	−0.18
GCP 4	*−1.74*	−0.52	−1.41	−1.35
GCP 5	*−1.37*	−1.01	−0.47	−0.52
GCP 6	*2.2*	1.74	1.55	1.18
GCP 7	1.88	*2.95*	1.80	1.79
GCP 8	*5.92*	4.51	2.01	3.34
GCP 9	1.14	*10.79*	0.73	0.55

Note: $Res = Ele_{GCPs} - Ele_{gen}$, where Ele_{GCPs} refers to the elevation of GCPs, Ele_{gen} refers to the elevation of generated DSMs. The best results are underlined, while the worst results are italicized.

TABLE 10 Quantitative results of AOI 1 and AOI 2.

	%						
	$[0, 1)$	$[1, 5)$	$[5, 10)$	$[10, \infty)$	**Per. $< T$ $T = 1$ m**	**RMSE (m)**	**ME (m)**
AOI 1 (*size*: 2437 × 2446 *pixels*; *resolution* = 0.8 *m*)							
SRDM	35.37	60.65	3.22	0.76	*35.37*	*3.13*	*1.82*
PSMNet	49.27	50.31	0.30	0.12	49.27	1.44	1.14
ACVNet	63.79	35.82	0.37	0.02	63.79	**1.25**	0.93
HF^2Net	64.42	35.21	0.34	0.03	64.42	1.44	0.93
AOI 2 (*size*: 8128 × 7051 *pixels*; *resolution* = 0.8 *m*)							
SRDM	45.34	47.38	5.69	1.59	*45.34*	*6.31*	*1.98*
PSMNet	52.80	44.78	1.92	0.50	52.80	2.39	1.29
ACVNet	63.75	33.90	1.83	0.52	63.75	2.69	1.14
HF^2Net	63.08	34.61	1.83	0.48	63.08	2.39	1.14

Note: T refers to the elevation residual threshold. % is the percentage within each elevation interval. The optimal results are shown in bold, while the worst results are italicized.

Performance assessment of largely flat regions

Experiments on AOI 1 and AOI 2 are used to observe the proposed workflow's performance on texture-less areas and repetitive regions, where the former features farmland areas and the latter features repetitive regions and concentrated residential areas. Table 10 displays the quantitative results. The elevation residuals threshold was set as $T = 1$ m for these two AOIs; that is, the points with elevation residuals larger than 1 m were viewed as false matching results and were filtered out from the accuracy statistics. From the quantitative aspect, all the DLSMs outperformed SRDM by a large margin. As described in Table 10, there were approximately 63% elevation residuals lower than 1 m for HF^2Net and ACVNet, while only 35.37% and 45.34% in the two AOIs for SRDM. Due to the severe false matching problem, the RMSE of SRDM were much larger than DLSMs. For example, the RMSE of SRDM were 3.13 and 6.31 m for the two AOIs, while the values of HF^2Net were only 1.44 and 2.39 m. The experimental results demonstrate the noteworthy superiority of the DLSMs in the texture-less areas and repetitive regions.

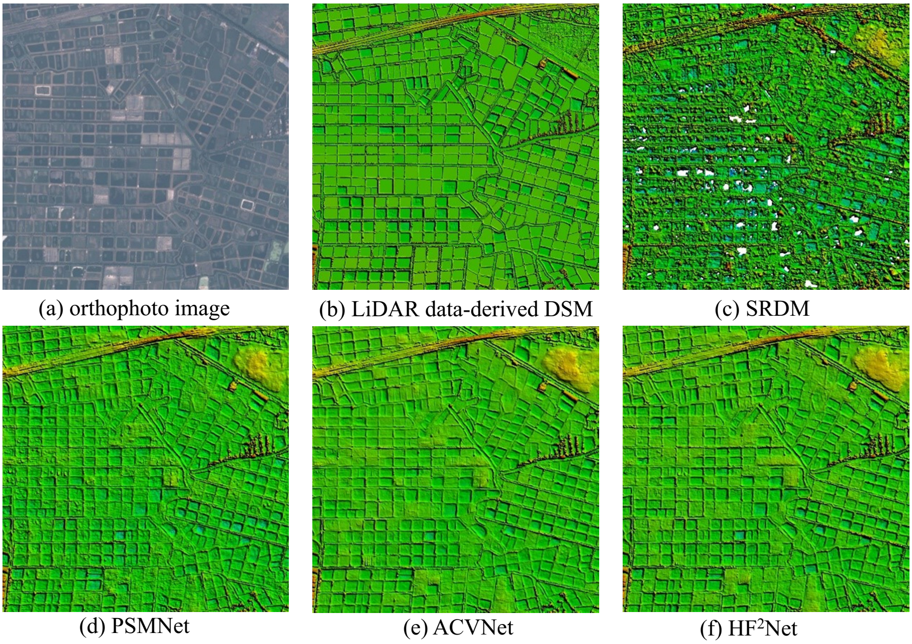


FIGURE 9 Visualization of the generated DSMs in AOI 1.

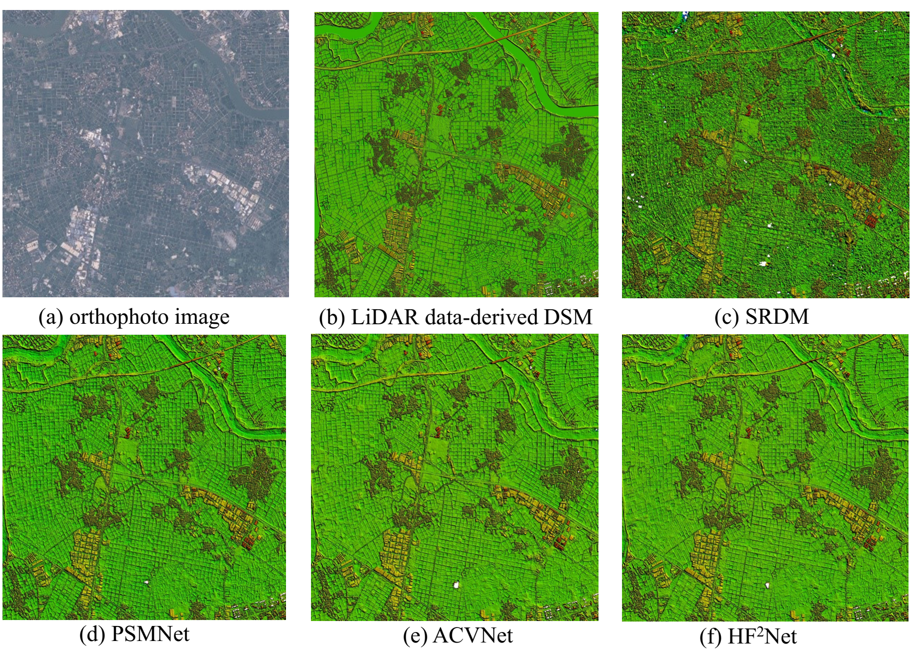


FIGURE 10 Visualization of the generated DSMs in AOI 2.

Among the three DLSMs, PSMNet performed worse than the other two methods. Nevertheless, its false matching rate was still lower than SRDM by 11.90% in AOI 1 and 7.46% in AOI 2. ACVNet has much better performance than PSMNet. Compared with ACVNet, HF^2Net achieved very similar elevation accuracy in the two AOIs. Referring to the statistical indicators, the proposed HF^2Net had nearly the same results as ACVNet.

TABLE 11 Quantitative results of AOI 4 and AOI 5.

	%						
	[0, 1)	[1, 5)	[5, 10)	[10, ∞)	Per. < T T = 5 m	RMSE (m)	ME (m)
AOI 4 (*size*: 4305 × 2831 *pixels*; *resolution* = 0.8 *m*)							
SRDM	28.53	53.62	10.84	7.01	82.15	7.54	3.49
PSMNet	32.41	38.61	18.32	10.66	*71.02*	7.94	*4.36*
ACVNet	37.40	45.82	9.73	7.05	83.22	*7.98*	3.48
HF^{2}Net	35.87	48.69	9.40	6.04	**84.56**	6.41	3.09
AOI 5 (*size*: 4330 × 2565 *pixels*; *resolution* = 0.8 *m*)							
SRDM	47.49	37.30	10.41	4.80	84.79	4.97	2.59
PSMNet	30.99	45.75	17.24	6.02	*76.74*	*5.46*	*3.42*
ACVNet	53.15	34.20	8.33	4.32	87.35	4.80	2.32
HF^{2}Net	53.85	34.59	7.99	3.57	88.44	**4.02**	2.13

Note: *T* refers to elevation residual threshold. Per. is the percentage within each elevation interval. (The optimal results are shown in bold, while the worst results are italicized).

To intuitively display the performance of HF^{2}Net, we visualized the generated DSMs of the two AOIs in Figures 9 and 10. As described in Figures 9 and 10, the generated DSMs of SRDM were disorderly, which is in contrast of the DLSMs. For example, the results of SRDM in the farmland areas of AOI 1 were unsmooth while HF^{2}Net and ACVNet almost completely reconstructed the entire region. In addition, all the DLSMs generally reconstructed the rivers (see the river at the top-right corner of Figure 9), while SRDM obtained much more severe false matching results in such areas. The results in Figures 9 and 10 verified the superiority of applying the proposed workflow in texture-less areas and repetitive regions.

Performance assessment on building parts

Experiments on AOI 4 and AOI 5 feature the building regions. AOI 4 focused on sparse and individual buildings, while AOI 5 focused on dense urban residential areas where occlusion and disparity discontinuity were more likely to happen. Table 11 list the accuracy results of AOI 4 and AOI 5. According to Table 11, the proposed HF^{2}Net again outperformed SRDM, while its superiority was not as significant as the results in texture-less areas and repetitive regions. For example, HF^{2}Net surpassed SRDM in elevation residual percentages by approximately 15% in the texture-less areas, while the difference was only about 2.5–4% in the building positions with the given elevation threshold. Note, however, that its superiority was increased to approximately 7% when the threshold was limited to 1 m, which indicates that the proposed HF^{2}Net obtained more elevation points with higher accuracy. Furthermore, HF^{2}Net decreased the RMSE of SRDM by about 1.13 m and 0.95 m in the two AOIs. It also yielded the highest accuracy among the three DLSMs, which showed its superiority. For example, HF^{2}Net outperformed the PSMNet and ACVNet in AOI 5 with an RMSE decrease of 0.78 and 1.44 m, respectively. Figures 11 and 12 illustrate the qualitative comparison of these methods in AOI 4 and AOI 5. As depicted in Figure 11, the building boundaries of SRDM were not as sharp as the results of HF^{2}Net. Meanwhile, fewer outliers seemed to occur around the buildings in the DSMs of HF^{2}Net than SRDM, which indicated that HF^{2}Net handled the occlusion and disparity discontinuous situations better.

Performance assessment on mountains

Experiments on AOI 3 and AOI 6 reflect the performance difference in mountains. As described in Table 12, the DLSMs did not show advantages compared with the conventional SRDM in the mountain regions. The quantitative accuracy of HF^{2}Net outperformed SRDM in AOI 3, but the situation was quite the contrary in AOI 6. It

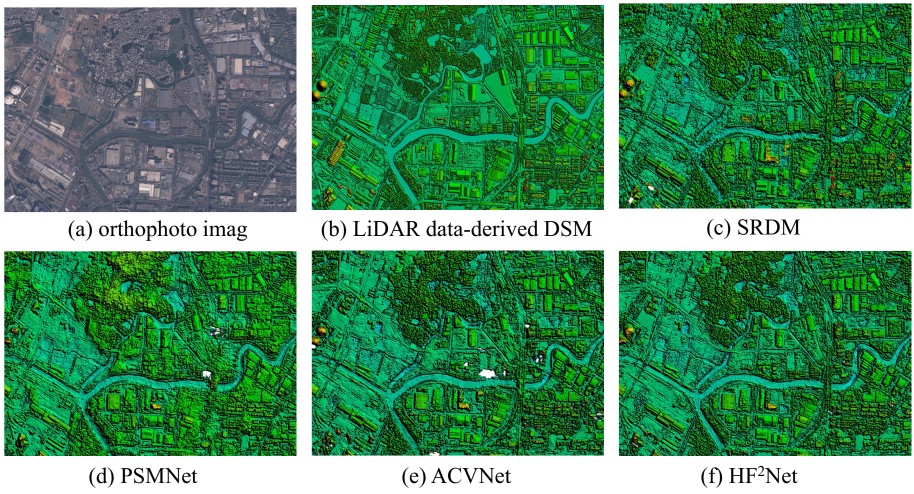


FIGURE 11 Visualization of the generated DSMs in AOI 4.

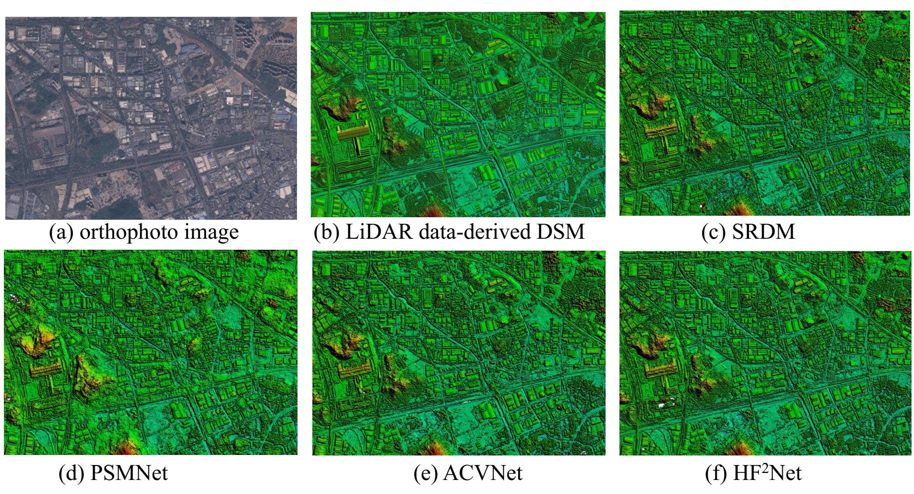


FIGURE 12 Visualization of the generated DSMs in AOI 5.

should be mentioned that the results of PSMNet were not as stable as the other methods in mountain regions. For example, it achieved a slightly higher accuracy than the other methods in AOI 3, while its results were largely worse than the other methods in AOI 6. In addition, though PSMNet had a higher valid residuals rate, its overall RMSE was similar to HF^{2}Net since more residuals of HF^{2}Net were less than 1 m. For the other two DLSMs, their performance was close to SRDM in the two AOIs.

Figures 13 and 14 display the generated DSMs of AOI 3 and AOI 6. It was difficult to observe obvious differences in the mountain regions among these methods. It should be noted that AOI 3 contained a large portion of concentrated residential areas (see the left-top and right-bottom corners of Figure 13), which indicates that the actual accuracy of the DLSMs in the mountain regions of AOI 3 may be lower than the values reported in Table 12.

3D visualization of two regions

In Figure 15, we display two large-scale scenes of the generated DSMs from a 3D perspective. Figure 15a depicts the result of largely texture-less area and repetitive regions; Figure 15b shows the DSM in residential areas and

TABLE 12 Quantitative results of AOI 3 and AOI 6.

	%						
	[0, 1)	[1, 5)	[5, 10)	[10, ∞)	Per. < T T = 10 m	RMSE (m)	ME (m)
AOI 3 (*size*: 4908 × 4664 *pixels*; *resolution* = 0.8 *m*)							
SRDM	30.37	49.60	14.19	5.84	*94.16*	*8.13*	*3.79*
PSMNet	36.08	52.10	9.27	2.55	97.45	5.93	2.87
ACVNet	45.49	42.38	8.64	3.49	96.51	5.46	2.51
HF^{2}Net	46.18	42.50	8.00	3.34	96.68	5.33	2.44
AOI 6 (*size*: 9295 × 9161 *pixels*; *resolution* = 0.8 *m*)							
SRDM	24.75	46.91	18.68	9.66	90.34	6.93	4.23
PSMNet	10.96	28.73	22.29	38.02	*61.99*	*28.94*	*13.09*
ACVNet	24.66	48.04	17.12	10.18	89.83	9.83	4.68
HF^{2}Net	23.64	48.01	17.53	10.82	89.18	9.77	4.79

Note: *T* refers to elevation residual threshold. *Per.* is the percentage within each elevation interval. The optimal results are shown in bold, while the worst results are italicized.

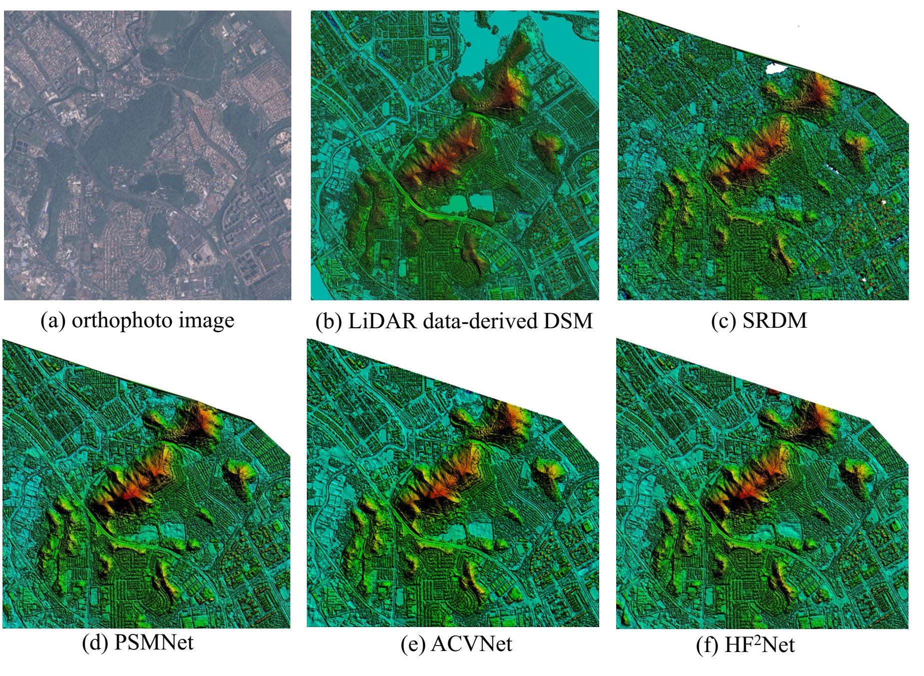


FIGURE 13 Visualization of the generated DSMs in AOI 3 (since the forward image did not cover the top-right corner of this AOI, the generated DSMs are not as complete as the LiDAR data-derived DSM).

mountains. It should be noted that the two scenes are the automatically reconstructed results of HF^{2}Net, except the main river in Figure 15a. As shown in Figure 15a, though the scene covers large areas of texture-less and receptive regions, most of the scene is well-matched, and the objects are clearly distinguished, such as bridges, buildings, and farmland. Figure 15b shows that the buildings reconstructed by the proposed HF^{2}Net have sharp boundaries and are neatly arranged. The buildings in dense residential areas, even the small and dense village-in-city, are also reconstructed. The mountains are reconstructed with apparent details. In addition, the narrow rivers in Figure 15b can be correctly matched by HF^{2}Net, which contrasts with the result of the main river in Figure 15a.

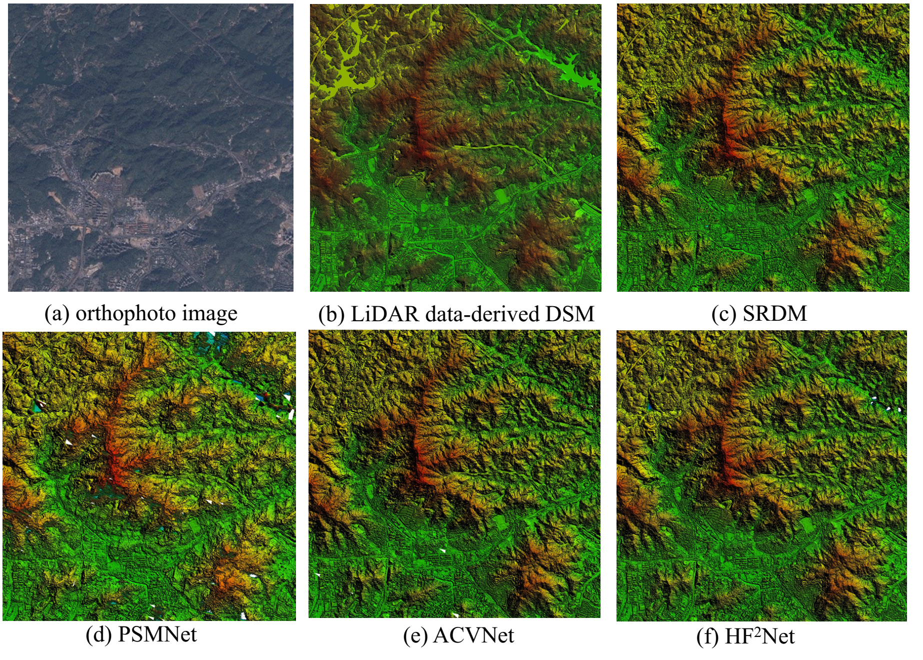


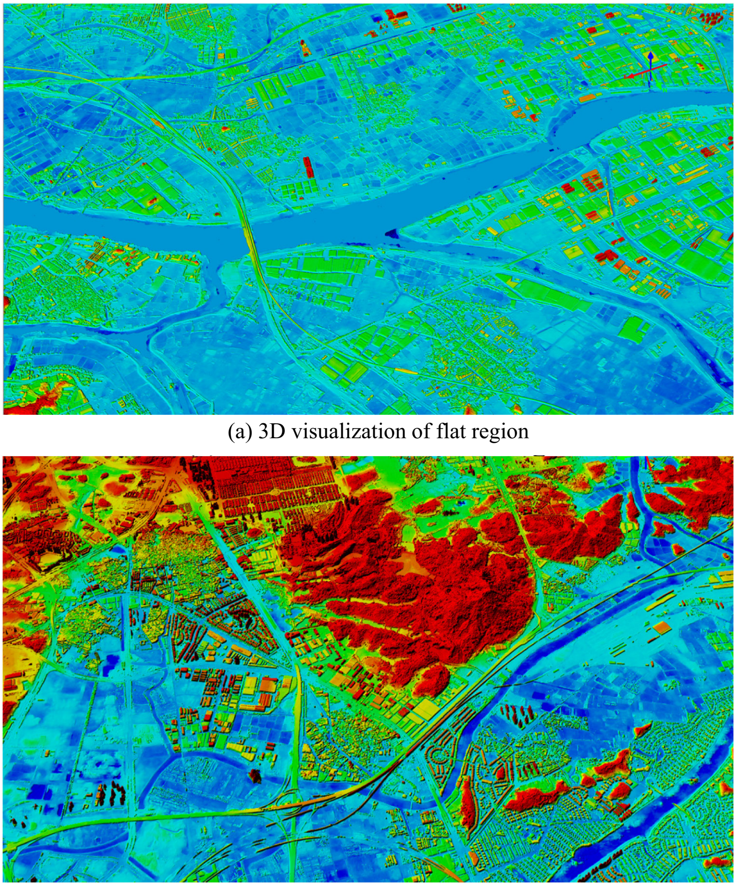
FIGURE 14 Visualization of the generated DSMs in AOI 6.

In a word, the reconstruction results of the two large-scale scenes showed the proposed DSM generation workflow's effectiveness and practical capability and evaluated its application potential.

DISCUSSION

Along with the experimental results of the two GF-7 stereos, we observed that the overall elevation accuracy of the proposed HF^{2}Net was higher than the conventional methods in most terrains. The result indicates that the DLSMs-based workflow has potential to recover more precise terrain and break through the accuracy bottleneck of the existing solution (Figures 9–12). Referring to different terrain types, the main features of the DLSMs-based workflow lie in orderly and precise recovery in intractable regions, regular and sharper building boundaries, and better overall accuracy; however, it does not perform apparent superiority over the conventional methods in mountain areas. Among the different DLSMs, our experiments indicate that the accuracy difference among their generated DSMs is very slight in most scenes. For example, though there is an apparent performance gap between PSMNet and ACVNet on the KITTI benchmark, their generated DSMs seem very close in most cases. Furthermore, we observed that the main difference among different DLSMs is their robustness to satellite stereos. For instance, the results of PSMNet on different terrains are significantly changed, which is less stable than HF^{2}Net and ACVNet. We also observed that the common problems of the DLSMs is the detail loss in their generated DSMs, such as small trees and brush areas, caused by the low-resolution disparity estimation. These findings suggest that follow-up research should investigate DLSMs' robustness and detail preservation more in-depth to enhance their practical capability.

Although the proposed DLSMs-based workflow demonstrates prominent superiority over conventional methods, it still has some shortcomings that must be fixed. The first problem is that more mismatching occurs on some tall buildings and wide rivers (i.e., the main river in Figure 15a). According to the initial results of Figure 15, we observed that the narrow rivers were well reconstructed while the wide rivers were not. Thus, we conducted



(a) 3D visualization of flat region

(b) 3D visualization of concentrated residentials and mountains

FIGURE 15 3D visualization of the generated DSMs with our DLSM-based workflow.

extra experiments on buildings with different heights to see whether the situation was highly correlated to long-range pixels. Experimental results verified our assumption, that is, HF^{2}Net achieved the best results on low-height buildings, and the reconstructed performance gradually deteriorated with buildings' height. The second problem is the unstable reconstruction in mountain areas. According to our experiments, we found that all the DLSMs caused inexplicable mismatches in some mountain areas, though the regions do not show significant differences from the surrounding scenes. By comprehensively analysing the DLSMs structure and the dataset's distribution on different terrains, we concluded that the main reason for this problem is the uneven distribution of the provided dataset, where many samples in mountain areas were filtered out. Thus, the terrain features of mountains were not well learned by the DLSMs.

The above analysis suggests that the proposed HF^{2}Net should be further investigated to enhance its capability of matching long-range pixels. In addition, the dataset's distribution should be carefully considered when applying the proposed DLSMs-based workflow for large-scale DSM generation.

CONCLUSIONS

DLSMs have shown superior performance on benchmark datasets while still facing practical problems when applying to satellite stereos. This paper develops a novel DLSM-based workflow for large-scale DSM generation from satellite stereos. The workflow includes pre-processing, disparity estimation and post-processing steps. The pre-processing step alleviates the problem of unmatched disparity range between satellite stereos and DLSMs and thus enables the application of DLSMs. The disparity estimation step provides a novel HF^{2}Net to enhance the overall disparity estimation accuracy and robustness. In detail, HF^{2}Net designs a hybrid feature extractor and a multi-scale cost filter. The hybrid feature extractor differentiates structural-context features and thus benefits the matching performance on intractable regions. The multi-scale cost filter filters out most matching errors and ensures accurate disparity estimation. The post-processing step generates initial DSM patches with estimated disparity maps and then refines them to obtain the final large-scale DSMs. Combining the three steps, we establish a complete DSM generation workflow, whose effectiveness and superiority have been demonstrated on the public US3D dataset and two GF-7 stereos.

In the future, we will test more geographical scenes and multi-source satellite stereos to observe the robustness of HF^{2}Net further. We will also pay more attention to improving the domain adaptation capability of HF^{2}Net.

ACKNOWLEDGEMENTS

The authors are grateful to the anonymous reviewers for their valuable comments. They thank the Johns Hopkins University Applied Physics Laboratory and IARPA for providing the data used in this study, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee for organizing the Data Fusion Contest. The numerical calculations were performed on the supercomputing system in the Supercomputing Center of Wuhan University.

FUNDING INFORMATION

This work in this paper was supported in part by the National Natural Science Foundation of China (grant numbers 42030102, 42192583 and 42001406); the Major Special Projects of Guizhou (grant number [2022]001); and the Program of High-Resolution Images Surveying and Mapping Application System-Phase II (grant number 42-Y30B04-9001-19/21).

CONFLICT OF INTEREST STATEMENT

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.

ORCID

Zhi Zheng https://orcid.org/0009-0004-8746-0591

Yongjun Zhang https://orcid.org/0000-0001-9845-4251

REFERENCES

Albanwan, H. & Qin, R. (2022) A comparative study on deep-learning methods for dense image matching of multi-angle and multi-date remote sensing stereo-images. *The Photogrammetric Record*, 37(180), 385–409. Available from: https://doi.org/10.1111/phor.12435

Bosch, M., Foster, K., Christie, G., Wang, S. & Brown, M. (2019) Semantic stereo for incidental satellite images. 2019 IEEE winter conference on applications of computer vision https://doi.org/10.1109/WACV.2019.00167

Bosch, M., Kurtz, Z., Hagstrom, S. & Brown, M. (2016) A multiple view stereo benchmark for satellite imagery. 2016 IEEE applied imagery pattern recognition workshop https://doi.org/10.1109/AIPR.2016.8010543

Chang, J.-R. & Chen, Y.-S. (2018) Pyramid stereo matching network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition https://doi.org/10.48550/arXiv.1803.08669

Chen, B., Qin, R., Huang, X., Song, S. & Lu, X. (2019) A comparison of stereo-matching cost between convolutional neural network and census for satellite images. arXiv preprint arXiv:1905.09147 https://doi.org/10.48550/arXiv.1905.09147

Cournet, M., Sarrazin, E., Dumas, L., Michel, J. & Fardet, Q. (2020) Ground truth generation and disparity estimation for optical satellite imagery. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 127–134. Available from: https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-127-2020

Facciolo, G., De Franchis, C. & Meinhardt-Llopis, E. (2017) Automatic 3D reconstruction from multi-date satellite images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops https://doi.org/10.1109/CVPRW.2017.198

Gao, J., Liu, J. & Ji, S. (2021) Rational polynomial camera model warping for deep learning based satellite multi-view stereo matching. Proceedings of the IEEE/CVF International Conference on Computer Vision https://doi.org/10.1109/ICCV48922.2021.00609

Gao, J., Liu, J. & Ji, S. (2023) A general deep learning based framework for 3D reconstruction from multi-view stereo satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195, 446–461. Available from: https://doi.org/10.1016/j.isprsjprs.2022.12.012

Gruen, A., Huang, X., Qin, R., Du, T., Fang, W., Boavida, J. et al. (2013) Joint processing of UAV imagery and terrestrial mobile mapping system data for very high resolution city modeling. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40, 175–182. Available from: https://doi.org/10.5194/isprsarchives-XL-1-W2-175-2013

Guo, X., Yang, K., Yang, W., Wang, X. & Li, H. (2019) Group-wise correlation stereo network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition https://doi.org/10.1109/CVPR.2019.00339

Han, Y., Qin, R. & Huang, X. (2020) Assessment of dense image matchers for digital surface model generation using airborne and spaceborne images–an update. *The Photogrammetric Record*, 35(169), 58–80. Available from: https://doi.org/10.1111/phor.12310

He, K., Zhang, X., Ren, S. & Sun, J. (2016) Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition https://doi.org/10.1109/CVPR.2016.90

He, S., Li, S., Jiang, S. & Jiang, W. (2022) HMSM-Net: hierarchical multi-scale matching network for disparity estimation of high-resolution satellite stereo images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 188, 314–330. Available from: https://doi.org/10.1016/j.isprsjprs.2022.04.020

Hirschmuller, H. (2007) Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2), 328–341. Available from: https://doi.org/10.1109/TPAMI.2007.1166

Hou, Y., Peng, J., Hu, Z., Tao, P. & Shan, J. (2018) Planarity constrained multi-view depth map reconstruction for urban scenes. *ISPRS Journal of Photogrammetry and Remote Sensing*, 139, 133–145. Available from: https://doi.org/10.1016/j.isprsjprs.2018.03.003

Hu, J., Xia, G.-S. & Sun, H. (2019) An SRTM-aided epipolar resampling method for multi-source high-resolution satellite stereo observation. *Remote Sensing*, 11(6), 678. Available from: https://doi.org/10.3390/rs11060678

Huang, X. & Qin, R. (2020) Post-filtering with surface orientation constraints for stereo dense image matching. *The Photogrammetric Record*, 35(171), 375–401. Available from: https://doi.org/10.1111/phor.12333

Huang, X., Qin, R., Xiao, C. & Lu, X. (2018) Super resolution of laser range data based on image-guided fusion and dense matching. *ISPRS Journal of Photogrammetry and Remote Sensing*, 144, 105–118. Available from: https://doi.org/10.1016/j.isprsjprs.2018.07.001

Huang, X., Zhang, Y.J. & Yue, Z.X. (2016) Image-guided non-local dense matching with three-steps optimization. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3(3), 67–74. Available from: https://doi.org/10.5194/isprsannals-III-3-67-2016

Ji, S., Liu, J. & Lu, M. (2019) CNN-based dense image matching for aerial remote sensing images. *Photogrammetric Engineering & Remote Sensing*, 85(6), 415–424. Available from: https://doi.org/10.14358/PERS.85.6.415

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A. et al. (2017) End-to-end learning of geometry and context for deep stereo regression. Proceedings of the IEEE International Conference on Computer Vision https://doi.org/10.1109/ICCV.2017.17

Le Saux, B., Yokoya, N., Hansch, R., Brown, M. & Hager, G. (2019) 2019 data fusion contest. *IEEE Geoscience and Remote Sensing Magazine*, 7(1), 103–105. Available from: https://doi.org/10.1109/MGRS.2019.2893783

Leotta, M.J., Long, C., Jacquet, B., Zins, M., Lipsa, D., Shan, J. et al. (2019) Urban semantic 3D reconstruction from multiview satellite imagery. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops https://doi.org/10.1109/CVPRW.2019.00186

Liu, X., Zhu, X., Zhang, Y., Wang, S. & Jia, C. (2023) Generation of concise 3D building model from dense meshes by extracting and completing planar primitives. *The Photogrammetric Record*, 38(181), 22–46. Available from: https://doi.org/10.1111/phor.12438

Lv, Z., Huang, H., Li, X., Zhao, M., Benediktsson, J.A., Sun, W. et al. (2022) Land cover change detection with heterogeneous remote sensing images: review, progress, and perspective. Proceedings of the IEEE https://doi.org/10.1109/JPROC.2022.3219376

Lv, Z., Zhong, P., Wang, W., You, Z. & Falco, N. (2023) Multi-scale attention network guided with change gradient image for land cover change detection using remote sensing images. IEEE Geoscience and Remote Sensing Letters https://doi.org/10.1109/LGRS.2023.3267879

Lv, Z., Zhong, P., Wang, W., You, Z. & Shi, C. (2023) Novel piecewise distance based on adaptive region key-points extraction for LCCD with VHR remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 5607709. Available from: https://doi.org/10.1109/TGRS.2023.3268038
Michel, J., Sarrazin, E., Youssefi, D., Cournet, M., Buffe, F., Delvit, J. et al. (2020) A new satellite imagery stereo pipeline designed for scalability, robustness and performance. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2, 171–178. Available from: https://doi.org/10.5194/isprs-annals-V-2-2020-171-2020
Pekel, J.-F., Cottam, A., Gorelick, N. & Belward, A.S. (2016) High-resolution mapping of global surface water and its long-term changes. *Nature*, 540(7633), 418–422. Available from: https://doi.org/10.1038/nature20584
Qin, R. (2019a) A critical analysis of satellite stereo pairs for digital surface model generation and a matching quality prediction model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 154, 139–150. Available from: https://doi.org/10.1016/j.isprsjprs.2019.06.005
Qin, R. (2019b) An operational pipeline for generating digital surface models from multi-stereo satellite images for remote sensing applications. 2019 IEEE International Geoscience and Remote Sensing Symposium https://doi.org/10.1109/IGARSS.2019.8897962
Qin, R., Huang, X., Liu, W. & Xiao, C. (2019) Pairwise stereo image disparity and semantics estimation with the combination of U-net and pyramid stereo matching network. 2019 IEEE International Geoscience and Remote Sensing Symposium https://doi.org/10.1109/IGARSS.2019.8900262
Rao, Z., He, M., Zhu, Z., Dai, Y. & He, R. (2020) Bidirectional guided attention network for 3-D semantic detection of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7), 6138–6153. Available from: https://doi.org/10.1109/TGRS.2020.3029527
Scharstein, D. & Szeliski, R. (2002) A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1), 7–42. Available from: https://doi.org/10.1109/SMBV.2001.988771
Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M. et al. (2017) A multi-view stereo benchmark with high-resolution images and multi-camera videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition https://doi.org/10.1109/CVPR.2017.272
Shen, H., Jiang, Y., Li, T., Cheng, Q., Zeng, C. & Zhang, L. (2020) Deep learning-based air temperature mapping by fusing remote sensing, station, simulation and socioeconomic data. *Remote Sensing of Environment*, 240, 111692. Available from: https://doi.org/10.1016/j.rse.2020.111692
Shen, Z., Dai, Y. & Rao, Z. (2021) Cfnet: cascade and fused cost volume for robust stereo matching. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition https://doi.org/10.1109/CVPR46437.2021.01369
Tao, R., Xiang, Y. & You, H. (2020) Stereo matching of VHR remote sensing images via bidirectional pyramid network. IGARSS 2020–2020 IEEE International Geoscience and Remote Sensing Symposium https://doi.org/10.1109/IGARSS39084.2020.9324093
Xie, J., Huang, G., Liu, R., Zhao, C., Dai, J., Jin, T. et al. (2020) Design and data processing of China's first spaceborne laser altimeter system for earth observation: GaoFen-7. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 1034–1044. Available from: https://doi.org/10.1109/JSTARS.2020.2977935
Xu, G., Cheng, J., Guo, P. & Yang, X. (2022) ACVNet: attention concatenation volume for accurate and efficient stereo matching. arXiv preprint arXiv:2203.02146 https://doi.org/10.48550/arXiv.2203.02146
Youssefi, D., Michel, J., Sarrazin, E., Buffe, F., Cournet, M., Delvit, J.-M. et al. (2020) CARS: a photogrammetry pipeline using Dask graphs to construct a global 3D model. IGARSS 2020–2020 IEEE International Geoscience and Remote Sensing Symposium https://doi.org/10.1109/IGARSS39084.2020.9324020
Zbontar, J. & LeCun, Y. (2015) Computing the stereo matching cost with a convolutional neural network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition https://doi.org/10.1109/CVPR.2015.7298767
Zbontar, J. & LeCun, Y. (2016) Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1), 2287–2318. Available from: https://doi.org/10.5555/2946645.2946710
Zhang, C., Cui, Y., Zhu, Z., Jiang, S. & Jiang, W. (2022) Building height extraction from GF-7 satellite images based on roof contour constrained stereo matching. *Remote Sensing*, 14(7), 1566. Available from: https://doi.org/10.3390/rs14071566
Zhang, F. & Wah, B.W. (2017) Fundamental principles on learning new features for effective dense matching. *IEEE Transactions on Image Processing*, 27(2), 822–836. Available from: https://doi.org/10.1109/TIP.2017.2752370
Zhang, K., Snavely, N. & Sun, J. (2019) Leveraging vision reconstruction pipelines for satellite imagery. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops https://doi.org/10.1109/ICCVW.2019.00269
Zhang, Y., Zhang, Y., Mo, D., Zhang, Y. & Li, X. (2017) Direct digital surface model generation by semi-global vertical line locus matching. *Remote Sensing*, 9(3), 214. Available from: https://doi.org/10.3390/rs9030214

Zhang, Y., Zou, S., Liu, X., Huang, X., Wan, Y. & Yao, Y. (2022) LiDAR-guided stereo matching with a spatial consistency constraint. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183, 164–177. Available from: https://doi.org/10.1016/j.isprsjprs.2021.11.003

Zhao, X., Zhou, Q., Dong, J. & Duan, Y. (2022) Digital elevation model-assisted aerial triangulation method on an unmanned aerial vehicle sweeping camera system. *The Photogrammetric Record*, 37(178), 208–227. Available from: https://doi.org/10.1111/phor.12419

How to cite this article: Zheng, Z., Wan, Y., Zhang, Y., Hu, Z., Wei, D., Yao, Y. et al. (2024) Digital surface model generation from high-resolution satellite stereos based on hybrid feature fusion network. *The Photogrammetric Record*, 39, 36–66. Available from: https://doi.org/10.1111/phor.12471