ISPRS Journal of Photogrammetry and Remote Sensing 207 (2024) 27–42

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Edge aware depth inference for large-scale aerial building multi-view stereo ☆

Song Zhang [a,b,c,d], ZhiWei Wei [a,b], WenJia Xu [e,*], LiLi Zhang [a,b], Yang Wang [a,b], JinMing Zhang [a,b], JunYi Liu [a,b]

[a] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[b] Key Laboratory of Network Information System Technology(NIST), Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China
[c] University of Chinese Academy of Sciences, Beijing 100190, China
[d] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China
[e] Beijing University of Posts and Telecommunications, Beijing 100190, China

ARTICLE INFO

Keywords:
Multi-view stereo
Aerial images
Building depth estimation
Building edge extraction
Building edge incorporation

ABSTRACT

Aerial building depth estimation is a crucial task in 3D digital urban reconstruction and learning-based multi-view stereo (MVS) methods have recently shown promising results in this field. However, these methods are mainly developed by modifying the general learning-based MVS framework for aerial depth estimation, which lack consideration about the intrinsic structures of buildings and result in insufficient accuracy. Therefore, we propose an end-to-end edge aware depth inference network for large-scale aerial building multi-views stereo, called **EG-MVSNet**, which incorporates the building edge information and jointly estimate the depth map and edge map. Firstly, we propose a novel Edge-Sensitive Network based on the differentiable Dynamic Sobel Kernels to obtain reliable building edge features while eliminating other irrelevant features. We further propose an UNet-like Edge Prediction Branch and a Building Edge-Depth Loss to constrain the model focus primarily on the building edge features. Notably, the pseudo ground truth (GT) edge map for each aerial image is obtained with classical gradient operators which do not require additional annotation. Secondly, to incorporate the edge features into the depth prediction module, we introduce an Inter-volume Adaptive Fusion Module that adaptively incorporates the edge features volume into a standard cost volume and guides the regularization of the cost volume. An Edge Depth Refinement Module is further proposed to performs 2D-guidance refinement and avoid over-smoothed or blurred depth boundaries. Extensive experiments on the WHU dataset and LuoJia-MVS dataset show that our model significantly outperforms state-of-the-art performance by more than 22% mean absolute error (MAE) compared to RED-Net and 57% MAE compared to MVSNet. Additionally, to validate our proposed model, we reconstruct a synthetic aerial building benchmark based on WHU dataset. The results as far as correctness and accuracy exceeded the results of other MVS methods in a between-method comparison by at least 12% in MAE metric. The dataset and code can be available at https://github.com/zs670980918/EG-MVSNet.

1. Introduction

Large-scale and highly accurate 3D reconstruction of the earth's surface is one of the important components of the metaverse. 3D urban reconstruction plays a critical role in monitoring human settlements and urban environments, evaluating disasters, and maintaining geographical information (Rottensteiner et al., 2014; Yu et al., 2021b; Xu et al., 2023). Stereo photogrammetry (Anon, 2015; Zhu et al., 2015) and marching cubes (Rothermel et al., 2012) have been the dominant classical algorithms for achieving large-scale and high-precision 3D urban reconstruction in recent decades. However, classical algorithms are vulnerable to mis-matching when dealing with non-Lambertian surfaces (e.g., river surface) or weak texture surfaces (e.g., gallery floor) (Aanæs et al., 2016), which may require significant manual effort in post-processing (Yu et al., 2021b).

In recent years, the MVSNet methods (Yao et al., 2018; Xiang et al., 2020; Yu et al., 2021a) have shown remarkable performance in estimating depth maps from multi-view images by constructing cost volumes based on CNN features and using 3D CNNs to regularize them. Some researchers have attempted to apply these methods to large-scale aerial MVS reconstruction based on aerial images, such as RED-Net (Liu and

☆ This work has been partially funded by the Fundamental Research Funds for the Central Universities with number 2023RC61, and the National Natural Science Foundation of China under Grant 62301063.

* Corresponding author at: Beijing University of Posts and Telecommunications, Beijing 100190, China.
E-mail address: xuwenjia@bupt.edu.cn (W. Xu).

https://doi.org/10.1016/j.isprsjprs.2023.11.020
Received 7 June 2023; Received in revised form 22 November 2023; Accepted 23 November 2023
Available online 28 November 2023

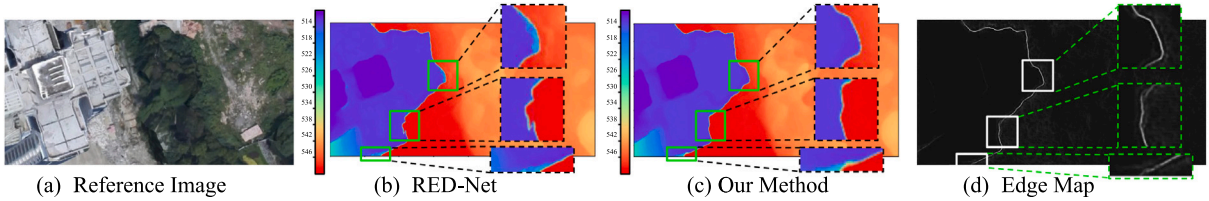


Fig. 1. Comparisons of the RED-Net (Liu and Ji, 2020) and our method in aerial building image depth estimation results. The black dashed boxes in the figure are zoomed-in views of the local details respectively. The green dotted boxes are the corresponding edge map detail views. The scale of colorbar represents the depth value. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Ji, 2020), MS-REDNet (Yu et al., 2021b), and HDC-MVSNet (Li et al., 2023a). However, these methods mainly modify the general learning-based MVS framework without considering the characteristics of the building objects themselves, resulting in low building reconstruction accuracy. Specifically, the aerial perspective often depicts buildings in close proximity to adjacent terrains. Therefore, when aerial images are taken with oblique photography, insufficient illumination, and the image resolution is low, it is difficult to distinguish the edges of buildings and terrains from the captured images, resulting in the edges of adjacent terrains interfering building surface depth value distribution (defined as **depth adhesion**). This effect engenders inaccuracies in aerial building depth estimation and may further impede accurate building reconstruction (Wei et al., 2022). As shown in Fig. 1(b), we can clearly observe that the previous method without considering the building edge information is prone to occur the depth adhesion between the foreground (buildings) and background (terrains) in the aerial building image depth estimation results.

In the edge detection domain, it is well known that edge information explicitly corresponds to drastic gradient changes (Li et al., 2021; Liu et al., 2018), and integrating the edge cues to guide the model prediction can help obtain more accurate and sharper masks and alleviate the problem of depth adhesion. For example, more accurate results are obtained after introducing the edge feature information, as shown in Fig. 1(c) and (d). Therefore, incorporate the edge information into multi-view stereo aerial building depth estimation would also be helpful to alleviate the problem of depth adhesion, but several challenges need to be solved first: (a) **How to effectively encode the building edge features**: Aerial images often have complex boundaries due to small fragments, shadows, occlusions, different backgrounds, and complex textures. But most of the current edge detection networks (Xie and Tu, 2015; Yu et al., 2017) are designed for general purposes and tend to extract irrelevant edge features, including those from the surrounding terrain, in addition to the building edge features. Incorporating these irrelevant edge features will interfere the building reconstruction result. Therefore, it is challenging to encode the building edge information from the complex scene information alone. (b) **How to effectively incorporate the building edge features**: The MVSNet algorithm (Yao et al., 2018) relies on the 3-D cost volume built from 2-D image surface features to generate a depth map. However, due to the inconsistency between 3-D cost volume and 2-D features in data dimension, it is challenging to directly guide the regularization process of the 3-D cost volume by the 2-D edge features.(c) **How to effectively verify the proposed method**: To our best knowledge, there are a few open-access aerial MVS benchmarks, *e.g.* WHU-dataset (Liu and Ji, 2020), LuoJia-MVS dataset (Li et al., 2023a), München dataset (Haala, 2014). However, these datasets contain not only building scenes, but also many different complex terrain scenes and are therefore not suitable for evaluating our method.

Regards the **problem (a)**, it is natural to improve the edge extraction network to focus primarily on extracting building edge features while minimizing the extraction of other irrelevant features. Two strategies may be effective: First, buildings exhibit rigid structures and their edge features tend to be linear. Thus, if we incorporate a module that can extract linear features into the network, it would improve the ability to encode linear edge features, namely the building edge features. Second, since the dataset for MVS has provided ground truth depth maps, from which we can generate tentative pseudo ground truth edge maps. Therefore, the pseudo edge maps can be used to calculate loss together with the predicted edge map obtained by regressing edge features, this would ensure that our network primarily focuses on the building edge features. Regards the **problem (b)**, since 3-D cost volume in MVS is constructed by matching similarities between 2-D features at different spatial positions and depths in different views using the differentiable homography (Yao et al., 2018), we can construct a 3-D edge feature volume based on extracted 2-D edge features based on the same mechanism. This would allow us to incorporate 2-D edge features for implicitly guiding the regularization of the 3-D cost volume. Furthermore, since the goal of MVS is to estimate a 2-D depth map, the 2-D feature guidance for explicitly refining the 2-D depth map can also help achieve more comprehensive edge feature fusion. Regards the **problem (c)**, we consider constructing a building-only aerial MVS benchmark based on existing open-access benchmarks to evaluate the effectiveness of our proposed modules.

Motivated by above thoughts, we proposed an end-to-end EG-MVSNet which introduces edge cues into large-scale aerial building depth estimation to obtain better performance. Firstly, to obtain precise and reliable building edge features, we propose an Edge-Sensitive Network (ESNet) based on the differentiable Dynamic Sobel Kernels (DSK). Since the traditional Sobel operator is suitable for extracting linear features (Sobel et al., 1968), we take it as a prior template and replaces the values with learnable parameters to achieve the differentiable DSK. Furthermore, we propose an UNet-like Edge Prediction Branch and a Building Edge-Depth Loss to constrain our model focus primarily on the building edge features and eliminating the irrelevant features. Additionally, the pseudo ground truth (GT) edge map is generated by the classical gradient operators (*e.g.* Sobel operator (Sobel et al., 1968) and Laplace operator) based on the GT depth map. Secondly, to incorporate the edge features into our task, we introduce from two aspects, which are 3D-guidance and 2D-guidance. (a) 3D-guidance: We propose an attention-based Inter-volume Adaptive Fusion Module to adaptively incorporate the edge feature volume into the standard cost volume to guide the depth map inference. (b) 2D-guidance: We propose an Edge Depth Refinement Module to explicitly uses edge features from the 2-D feature level to refine the estimated depth map. Finally, to verify the effectiveness of our proposed model, we reconstruct a synthetic aerial building benchmark based on the WHU dataset, called Aerial Building MVS Dataset. It is constructed by manually selecting images from the WHU dataset (Liu and Ji, 2020) and post-annotation mask processing. Extensive experiments on several benchmark datasets (*e.g.* WHU dataset (Liu and Ji, 2020), LuoJia-MVS dataset (Li et al., 2023a), Aerial Building MVS Dataset) demonstrate that our approach achieves excellent performance and demonstrates competitive generalization ability compared to all-listed methods. Meanwhile, our method exhibits superior visualization for aerial building scene on our Aerial Building MVS Dataset, especially on edge regions.

In summary, the contributions of this paper are as follows:

- We propose an end-to-end network (EG-MVSNet) with collaborative optimization of edge map and depth map for large-scale aerial building depth estimation, which alleviates the problem of depth adhesion by incorporating the edge feature cues into the standard MVSNet pipeline.
- We propose an Edge-Sensitive Network based on differentiable Dynamic-Sobel Kernel to obtain precise and reliable while eliminating other irrelevant features. And we propose an UNet-like Edge Prediction Branch and a Building Edge-Depth Loss to constrain our model focus primarily on the building edge features.
- We propose an Inter-volume Adaptive Fusion Module to incorporate the edge feature volume into the standard cost volume to guide the depth map inference. And we also adopt an Edge Depth Refinement Module to refine the depth map by explicitly utilizing the edge features.

2. Related work

2.1. Close-range universal multi-view stereo

MVS has been exploited for decades with traditional methods such as COLMAP (Schonberger and Frahm, 2016) and ACMM (Xu and Tao, 2019) which achieve great and robust results. COLMAP (Schonberger and Frahm, 2016), as introduced by Schonberger et al. employs manual feature crafting and simultaneously estimates pixel-level view selection, depth map, and surface normals, leveraging both photometric and geometric priors. In a similar vein, Xu et al. propose ACMM (Xu and Tao, 2019), featuring multi-scale geometric consistency, adaptive checkerboard sampling, and multi-hypothesis joint view selection. Despite their remarkable achievements, traditional MVS methods rely on handcrafted features, rendering them unsuitable for scenarios involving non-Lambertian surfaces, regions with low textures, and texture-less areas where photometric consistency is less dependable. Rather than using traditional hand-crafted image features, CNN-based stereo methods (Yao et al., 2018; Xiang et al., 2020; Yu et al., 2021a) have been successfully applied to the close-range MVS task as a result of the success of deep learning-based stereo methods. Using 2D CNN to extract features, Yao et al. (2018) constructed a 3D cost volume using differentiable homography warping, followed by regularizing this cost volume with 3D CNN. It demonstrated significant improvements on many MVS benchmarks. PruMVS (Xiang et al., 2020) pioneers an approach that incorporates pruning techniques for 2D and 3D convolutional modules in MVS. It introduces a mixed backpropagation process to learn a pruning mask for 2D CNNs. Furthermore, PruMVS explores the performance of 3D CNNs and finds that even the smallest module maintains high performance. AACVP-MVSNet (Yu et al., 2021a) integrates self-attention layers for hierarchical feature extraction, enhancing the model's capability to capture crucial depth inference information at various levels. It also introduces a novel similarity measurement technique for aggregating pairwise costs, departing from conventional variance-based methods. Moreover, through a convolutional gated recurrent unit (GRU) (Cho et al., 2020) instead of 3D CNNs, R-MVSNet (Yao et al., 2019) regularized 2D cost maps sequentially across depths, which reduced memory usage and enabled high-resolution reconstruction. D2HC-RMVSNet (Yan et al., 2020) utilizes a hybrid architecture to achieve high-resolution reconstruction, DHU-LSTM, which combines the benefits of LSTM (Hochreiter and Schmidhuber, 1997) and U-Net (Ronneberger et al., 2015). The accomplishments attained by CNNs in close-range MVS tasks have fostered a surge of subsequent investigations and explorations within the realm of large-scale aerial MVS.

2.2. Large-scale aerial multi-view stereo

Despite advances in close-range object reconstruction, there are still challenges in applying the neural networks to aerial images because of the differences (*e.g.* scale range, scene type) between close-range and aerial MVS datasets. Therefore, in order to achieve accurate and effective multi-view stereo depth estimation in the aerial domain, some subsequent methods achieve superior aerial image depth estimation, e.g RED-Net (Liu and Ji, 2020), HDC-MVSNet (Li et al., 2023a), and MS-REDNet (Yu et al., 2021b). In contrast to R-MVSNet (Yao et al., 2019), which stacks three gated recurrent units (GRUs), RED-Net (Liu and Ji, 2020) was the first network created for MVS aerial image depth estimation. It has surpassed the traditional state-of-the-art MVS aerial image depth estimation techniques. In order to produce full-resolution depth with a wealth of contextual information, HDC-MVSNet (Li et al., 2023a) propose a hierarchical deformable cascade MVS network for aerial image depth estimation. This network simultaneously performs high-resolution multi-scale feature extraction and hierarchical cost volume module construction. Despite the above two methods have been proposed for large-scale aerial image depth estimation, neither of them specifically designs an exclusive strategy for urban reconstruction due to the large-scale aerial images containing very complex terrain and a wide variety of feature elements. Therefore, MS-REDNet (Yu et al., 2021b) combines the MVS and semantic segmentation methods to achieve high-precision and complete reconstruction of buildings. The multi-view stereo approach estimates the depth map, enabling the generation of digital surface model (DSM) and digital orthophoto map (DOM). Initial and assisted segmentation techniques are utilized to obtain the building segmentation map for building footprint extraction. Through this pipeline, the method generates Level of Detail 1 building models from multi-view aerial images. Nonetheless, MS-REDNet lacks a clear strategy for building depth estimation within the multi-view stereo process. It additionally utilizes semantic segmentation to extract the building segmentation map, establishing a sequential pipeline for automated building reconstruction. Motivated by this, our study primarily concentrates on enhancing the accuracy of building depth estimation in multi-view stereo. To this end, we innovatively combine edge detection and multi-view stereo, enabling collaborative optimization and incorporating the building edge information for achieving high-precision building depth estimation.

2.3. Edge cues in 3D domain

Edge information plays a key role in various computer vision problems (Yu et al., 2017), e.g., super-resolution (Xu et al., 2018b; Fang et al., 2020; Xu et al., 2018a), image classification (Xu et al., 2020; Duan et al., 2019; Xu et al., 2022), and 3D reconstruction (Li et al., 2023b; Qi et al., 2020). Recently, there have been experimental investigations in the field of 3D domain that explore the incorporation of edge information into their pipelines. EdgeStereo (Song et al., 2020) seamlessly incorporates edge information and regularization into the stereo disparity prediction network. The CP-RPN encompasses a context pyramid module for encoding multi-scale contextual information and a compact residual pyramid module for iterative refinement. Moreover, the edge sub-network is designed to effectively integrate edge information through feature embedding and an edge-aware smoothness loss, ensuring the preservation of fine details during the disparity estimation process. Furthermore, in the field of multi-view stereo (MVS), recent endeavors have emerged that harness edge cues to advance the precision of depth estimation. Notably, EPNet (Su and Tao, 2023) introduces a hierarchical edge-preserving residual learning framework, strategically designed to progressively alleviate upsampling inaccuracies. Furthermore, the incorporation of edge information enhances the refinement of multi-scale depth estimation. Concurrently, DDL-MVS (Ibrahimli et al., 2023) presents a comprehensive framework wherein depth map and edge map estimation are jointly pursued, with explicit utilization of

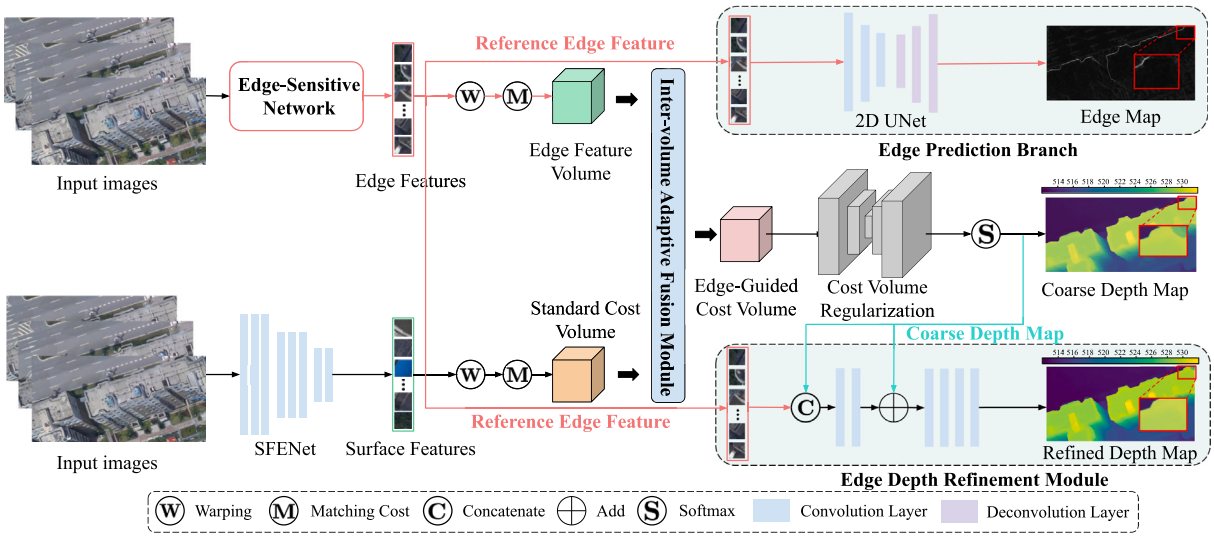


Fig. 2. Illustration of our EG-MVSNet. Our network consists of two parts, which can obtain the edge map, coarse depth map and refined depth map respectively (The detail comparison of each map are exhibited in the red boxes). The bold parts in the figure are our proposed modules. The architectures of SFENet and cost volume regularization network are derived from the RED-Net (Liu and Ji, 2020). Given the similar intermediate output with RED-Net (Liu and Ji, 2020), we refrain from elaborating on the specifics of SFENet and the cost volume regularization network within our methodology.

the edge map for subsequent depth map refinement. Motivated by these methodologies, we embark on a novel exploration involving the integration of edge cues within the MVS pipeline, aimed at advancing the precision of aerial building depth estimation. Distinctively, we have devised a task-specific module, denoted as ESNet, which takes into account the unique requisites of our task and the inherent architectural attributes of buildings. Furthermore, we have introduced an innovative manner for cost volume MVS-based edge feature 3D guidance, setting it apart from established methods.

3. Methodology

Our approach is in light of capturing and incorporating edge cues into the standard MVS (Yao et al., 2018) pipeline to alleviate the depth adhesion issue and achieve superior aerial building reconstruction quality. To address this, we introduce EG-MVSNet, an end-to-end trainable framework comprising two parts. These parts are dedicated to extracting and integrating edge feature information, thereby enhancing the accuracy of the depth estimation. Fig. 2 illustrates the architecture of our framework.

Part 1 (Edge feature extraction): Part 1 aims to obtain precise and reliable building edge features for later edge features incorporation (Part 2). Specifically, to effectively encode the edge features, we propose an Edge-Sensitive Network (ESNet) to extract edge features $\left\{F_{e_i}\right\}_{i=0}^{N-1}$ from N images $\left\{I_i\right\}_{i=0}^{N-1} \in \mathbb{R}^{H\times W\times 3}$. Then the differentiable homography warping (Yao et al., 2018) based on the depth hypothesis planes and the extracted edge features $\left\{F_{e_i}\right\}_{i=0}^{N-1}$ is used to construct an edge feature volume V_e which can be incorporate into the standard cost volume V for implicitly 3D-guiding the regularization in Part 2. Moreover, to ensure our ESNet primarily focus on building edge features rather than other irrelevant features, we utilize an UNet-like Edge Prediction Branch (EPB) to predict the edge map E and a Building Edge-Depth Loss (BED-Loss) to calculate the loss between the edge map E and pseudo ground truth edge map $\hat{E}$.

Part 2 (Edge feature incorporation): Part 2 aims to incorporate the building edge features (Part 1) into the standard MVS pipeline to guide the depth map inference from two aspect (2D-guidance and 3D-guidance). In particular, we firstly adopt the similar process as MVSNet (Yao et al., 2018) to construct the cost volume V. Then, to achieve the implicitly 3D-guidance, we propose an Inter-volume Adaptive Fusion Module (IAFM) to fuse the edge feature volume V_e and the standard cost volume V to obtain the edge-guided cost volume $\hat{V}$ to predict the coarse depth map D_c by 3D UNet. Furthermore, to achieve the explicitly 2D-guidance, we design an Edge Depth Refinement Module (EDRM) to refine coarse depth map D_c based on edge feature F_{e_0} to generate the refined depth map D_r.

The output of each part, *e.g.* Part 1 (edge map) and Part 2 (coarse depth map, refined depth map), undergoes individual loss calculations to impose constraints. As these parts interact and influence each other, our model achieves optimal results of each part through collaborative optimization, considering the respective losses. While standard MVS pipeline is similar to the MVSNet (Yao et al., 2018), we will elaborate on Part 1 (edge feature extraction) and Part 2 (edge feature incorporation) in subsequent sections to provide comprehensive insights.

3.1. Edge feature extraction

Part 1 aims to extract the reliable and precise 2-D building edge features, we propose two modules, which are Edge-Sensitive Network (edge feature encoder) and UNet-like Edge Prediction Branch (edge feature decoder) for guiding aerial building depth estimation in subsequent stages.

3.1.1. Edge-sensitive network (ESNet)

ESNet is proposed to accurately extract building edge features. Unlike general edge detection networks (Xie and Tu, 2015; Yu et al., 2017), which lack specific aptitude for identifying building edges in aerial imagery, ESNet is designed with the characteristics of buildings and MVSNet in mind. Specifically, to capture the rigid features of buildings, we transform the traditional Sobel operator (Sobel et al., 1968) into four different-directional Dynamic-Sobel Kernels (DSKs) which are learnable, since the Sobel operator is widely used for obtaining linear information in classical edge detection algorithms. Then a multi-directional linear feature perception strategy (Dynamic-Sobel Kernel Conv Layer in Fig. 3 right) is proposed to combine the different DSKs. Additionally, we adopt a network parameters-sharing strategy for multiple views to reduce computational complexity. Our ESNet, illustrated in Fig. 3, consists of three cascaded Dynamic-Sobel Kernel Conv Layers, each employing four different-directional Dynamic-Sobel Kernels to downsample the feature map and extract edge features. This mechanism ensures appropriately scaled features for the subsequent

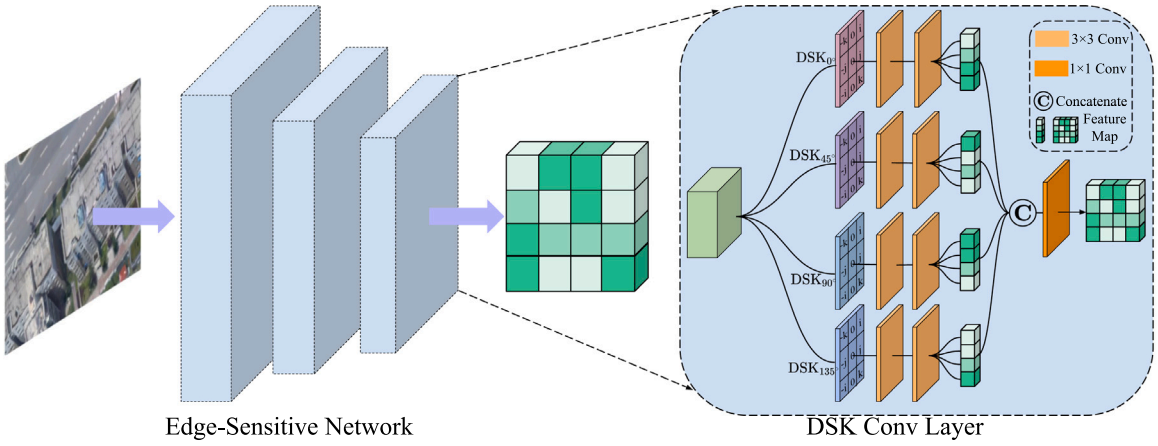


Fig. 3. Illustration of ESNet. Left: the ESNet consists of three cascades DSC Conv Layer. Right: the four-way DSKs for each DSC Conv Layer.

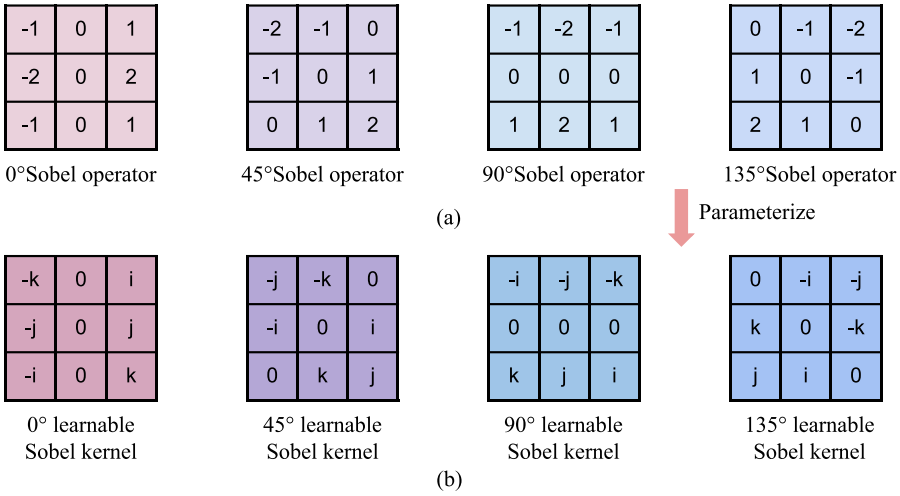


Fig. 4. Illustration of different forms of Sobel operator (Sobel et al., 1968) and learnable Sobel Kernel. (a) shows four different forms of unlearnable Sobel operators. (b) shows the differentiable Sobel Kernel after parameterization.

MVSNet process. The Dynamic-Sobel Kernels and the Dynamic-Sobel Kernel Conv Layer will be elaborated in detail in the following sections.

Differentiable Dynamic-Sobel Kernel (DSK): For aerial building depth estimation, the primary edges in the image are supposed to be the outlines of the building. Since the building is a rigid structure, most of the edge features are linear features. Thus, enhancing the network to extract linear features would improve the ability to encode linear edge features, namely the building edge features, and DSK is designed in light of this purpose. DSK transforms the traditional Sobel operation (Sobel et al., 1968) into a kernel due to the Sobel operator's ability to obtain linear information, as shown in Fig. 4(a). We can observe that the Sobel operator comprises two pairs of the most popular 3×3 kernels (0°-90° Sobel kernels and 45°-135° Sobel kernels (Sobel et al., 1968)). The maximum responses from these kernels are intended to be produced at gradient directions of 0°, 90°, 45° and 135°. For instance, 0°-90° Sobel kernels and 45°-135° Sobel kernels in grid R can be formulated as:

$$\begin{aligned} S_{0^\circ} &= \{-1,0,1,-2,0,2,-1,0,1\} \\ S_{45^\circ} &= \{-2,-1,0,-1,0,1,0,1,2\} \\ S_{90^\circ} &= \{-1,-2,-1,0,0,0,1,2,1\} \\ S_{135^\circ} &= \{0,-1,-2,1,0,-1,2,1,0\}. \end{aligned} \tag{1}$$

Since the Sobel operator (Sobel et al., 1968) is unlearnable, and its fitting ability is not powerful enough for depth estimation of aerial building images with a large number of data. Furthermore, considering the similarity between the structure of the Sobel operator and a convolutional kernel, we propose the conversion of the Sobel operator into a differentiable Sobel Kernel to enhance the ability for linear edge feature extraction. Therefore, to precisely capture the linear edge information, we present a Dynamic-Sobel Kernel, which uses the classical Sobel operator as a prior template and then parameterizes the fixed values with learnable parameters to produce the differentiable Sobel Kernel (as shown in Fig. 4(b)). The differentiable Dynamic-Sobel Kernel can be formulated as:

$$\begin{aligned} \mathrm{DSK}_{0^\circ} &= \{-k,0,i,-j,0,j,-i,0,k\} \\ \mathrm{DSK}_{45^\circ} &= \{-j,-k,0,-i,0,i,0,k,j\} \\ \mathrm{DSK}_{90^\circ} &= \{-i,-j,-k,0,0,0,k,j,i\} \\ \mathrm{DSK}_{135^\circ} &= \{0,-i,-j,k,0,-k,j,i,0\}, \end{aligned} \tag{2}$$

where the i, j, k represent the parameters of the convolutional kernel.

Dynamic-Sobel Kernel Conv Layer (DSK Conv Layer): Based on the above Dynamic-Sobel Kernels, we can convert the Sobel operator into a learnable convolutional kernel to participate in the optimization of the model, thus improving the capability for encoding linear features. However, due to the linear features are composed of horizontal, vertical and oblique lines, if we use single directional linear feature extraction DSK (*e.g.* 0° DSK or 90° DSK), the feature extraction of oblique lines may not be sufficient due to the structure of convolutional kernel. Thus, we combine the above different DSKs into a layer to encode linear

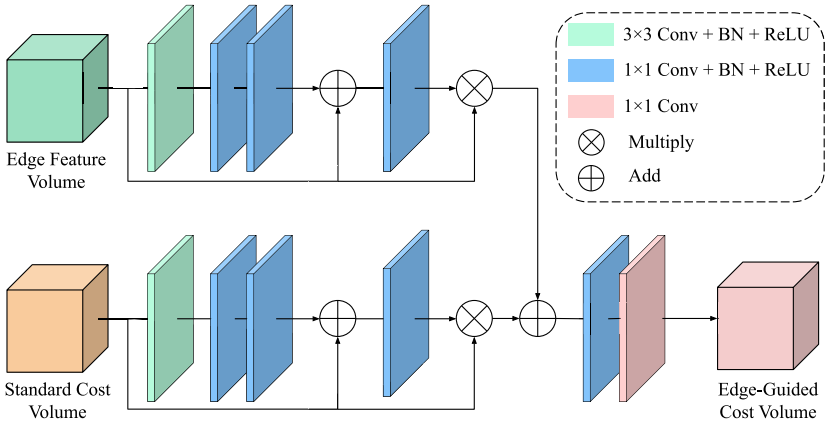


Fig. 5. Illustration of Inter-volume Adaptive Fusion Module.

features in any direction to effectively capture the edge structure of building and is implemented by Dynamic-Sobel Kernel Conv Layer, as shown in Fig. 4(b). The illustration of our DSK Conv layer is shown in Fig. 3, we use a four-way differentiable DSK to construct our DSK Conv Layer to perceive the complex building edges, and then concatenate the feature maps extracted from each branch to obtain the final edge feature map. Each branch corresponds to the 0° DSK, 45° DSK, 90° DSK, and 135° DSK, and each of them uses a different type of differentiable DSK to extract linear features. Two convolution layers are then applied for denoising and downsampling features. Finally, we fuse the extracted feature information from each branch to obtain the final building edge feature map. The formulation of our DSC Conv Layer is defined as (3):

$$\begin{aligned} f_{b1} &= w_{b_1} \odot \mathrm{DSK}_{0^\circ}(\hat{F}) \\ f_{b2} &= w_{b_2} \odot \mathrm{DSK}_{45^\circ}(\hat{F}) \\ f_{b3} &= w_{b_3} \odot \mathrm{DSK}_{90^\circ}(\hat{F}) \\ f_{b4} &= w_{b_4} \odot \mathrm{DSK}_{135^\circ}(\hat{F}) \\ F_e &= [f_{b_1}, f_{b_2}, f_{b_3}, f_{b_4}], \end{aligned} \tag{3}$$

where $[\cdot]$ represent the concatenate operation, w_{b_i} respectively represents the convolutional weights of each branch, $\odot$ represents the element-wise product.

3.1.2. UNet-like edge prediction branch (EPB)

As we mentioned above, ESNet has extracted the edge features for latter incorporation. Based on this, we propose an UNet-like Edge Prediction Branch to regress the edge features to obtain an edge map to participate in the loss calculation, which in turn constrains our ESNet to primarily focus on the edge features of buildings while minimizing other irrelevant features. Considering that edge map prediction is a semantic-like segmentation problem (Xie and Tu, 2015) and UNet (Ronneberger et al., 2015) has achieved great success in semantic segmentation, we consider designing our dedicated branch based on UNet that performs regression of edge features to generate an edge map for loss calculation. The architecture of our EPB is illustrated in the higher right part of Fig. 2. Specifically, the EPB uses a general 2D UNet network (Ronneberger et al., 2015) to regress the reference edge feature into an edge map. The EPB has similarities with the 2D UNet (Ronneberger et al., 2015), in that it employs an encoder–decoder structure to efficiently gather neighboring information from a wide receptive field, without incurring excessive memory and computational expenses. In the downsampling phase, edge features are progressively downsampled into four edge intermediate feature maps corresponding to different scales. Subsequently, the decoder performs upsampling of the edge intermediate feature maps to restore the original image size to obtain the edge map. Our EPB serves to explicitly leverage the edge features while also implicitly refining the depth map, thus mitigating the problem of depth adhesion. This prediction process allows for improved accuracy and reliability in the edge features, which play the key role in our overall framework. Moreover, the EPB also enables us to obtain visualized results, which can help us to better understand the performance of our proposed method. This visualization is critical for verifying the effectiveness of our approach and ensuring that it is meeting the intended objectives.

3.2. Edge feature incorporation

Part 2 aims to incorporate the 2-D building edge features into our task, we introduce from two aspects. (1)3D-guidance (Inter-volume Adaptive Fusion Module), incorporate the extracted 2-D building edge features in Part 1 into the standard MVS to guide the regularization of the 3-D cost volume and implemented by Inter-volume Adaptive Fusion Module, thus improve the accuracy of MVS; (2) 2D-guidance (Edge Depth Refinement Module), incorporate the edge features into the coarse depth map to obtain more detailed edge information and clarify the depth boundary and implemented by Edge Depth Refinement Module, thus improve the edge feature extraction. The two modules are jointly optimized and can achieve higher quality than either.

3.2.1. Inter-volume adaptive fusion module (IAFM)

The IAFM is designed to incorporate the extracted 2-D building edge features in Part 1 into the MVS to guide the regularization of the 3-D cost volume. Therefore, we need to transform the 2-D edge features into a 3-D features to implement the incorporation. The 3-D cost volume in MVS is constructed by matching similarities between 2-D feature points at different spatial positions in different views at different depths. Inspired by this mechanism, we also adopt the differentiable homography warping (Yao et al., 2018) to construct the 3-D edge feature volume. The formulation of differentiable homography warping is defined as (4):

$$H_i^{(d)} = d K_i T_i T_{ref}^{-1} K_{ref}^{-1}, \tag{4}$$

where T, K represent camera extrinsics and intrinsics respectively. Through this operation, we convert the 2D edge features to the 3D domain. However, not all edge matching information within the constructed edge feature volume is valid, and we aim to selectively fuse only the valid information into the cost volume during the fusion process, disregarding the irrelevant information. Directly summing the volumes may result in the invalid cost information from the edge feature volume affecting the valid cost information of the cost volume, such as surface cost information. Thus, if we can adaptively integrate the valid edge cost information into the cost volume, it would mitigate impact of invalid cost information. Benefited from the attention mechanism, we propose an attention-based Inter-volume Adaptive Fusion

Module (IFAM) that enhances the fusion of edge information and the cost volume by leveraging attention mechanisms (constructed by multiple stacked convolutions) to enhance the edge feature volume and suppressing non-edge cost information. The structure of our IFAM is illustrated in Fig. 5. Specifically, the IFAM calculates the edge-enhanced volume of the edge feature volume V_e by applying multiple stacked convolutions with skip connections. Similarly, an information-enhanced volume is obtained for the standard cost volume $\hat{V}$ using a similar mechanism. The IFAM then fuses these volumes by element-wise addition, resulting in the final edge-guided cost volume ($\hat{V}$) after two convolutional layers. Thus, the IFAM can be defined as Eq. (5).

$$\begin{aligned}\tilde{V}_e &= w_{12} \odot (w_{11} \odot V_e + V_e) \otimes V_e \\ \tilde{V} &= w_{22} \odot (w_{21} \odot V + V) \otimes V \\ \hat{V} &= w_{23} \odot (\tilde{V}_e + \tilde{V}),\end{aligned} \tag{5}$$

where $\odot$ indicates the Hadamard multiplication, $\otimes$ denotes the multiply operation, w_i represents the weights of the convolutional layers, $\tilde{V}_e$ and $\tilde{V}$ represent corresponding attention-enhanced intermediate volumes. In this way, our IAFM can obtain more accurate and effective depth maps by using edge features to guide our standard cost volume.

3.2.2. Edge depth refinement module (EDRM)

Although the depth map obtained from the probabilistic volume is a valid output, it may suffer from over-smoothed or blurred depth boundaries due to the involvement of large receptive fields in regularization, which is a common issue in semantic segmentation and image denoising tasks (Yao et al., 2018). Thus, if we incorporate the edge features into the coarse depth map to obtain more detailed edge information and clarify the depth boundary, it would alleviate above issue. Inspired by this, we address this challenge by exploring the use of 2D edge features to refine our initial coarse depth map, aiming to enhance the building edge depth estimation results and improve the overall quality of the depth map. To this end, we propose an Edge Depth Refinement Module (EDRM), specifically designed to optimize the depth map by leveraging the extracted reference edge feature. The reference for the EDRM is illustrated in the lower right part of Fig. 2. In this module, we first concatenate the coarse depth map D_c with the reference edge feature F_{e_0}, followed by incorporating the edge residual information obtained through two convolutional layers applied to the coarse depth map. This aggregation process enables the integration of edge information and yields the edge-enhanced depth map. Finally, we employ four additional convolutional layers to filter the edge-enhanced depth map and generate our refined depth map D_r. The formulation of the EDRM is defined as follows:

$$D_r = w_2 \odot ((w_1 \odot [F_{e_0}, D_c]) + D_c), \tag{6}$$

where $\odot$ indicates the Hadamard multiplication, w_i represents the weights of the convolutional layers.

3.3. Loss function

To obtain the high-quality outputs of each branch (Part 1: edge map, Part 2: depth map), we adopt two loss functions: Building Edge-Depth Loss (for edge map), Building Surface-Depth Loss (for depth map).

Building Edge-Depth Loss (BED-Loss): BED-Loss is designed to measure the difference in the edge map between the prediction and the pseudo ground truth to constrain our model (*e.g.* ESNet) focus primarily on the building edge features. However, since the Aerial Building MVS Dataset lacks the ground truth (GT) edge map, it is difficult to directly calculate the loss term by the depth map. Thus, we consider generating a pseudo GT edge map from the GT depth map $\hat{D}$ to participate in the loss calculation. Due to building boundaries are typically linear edges, which can be easily extracted using edge detection operators such as the Sobel operator (Sobel et al., 1968) and Laplace operator, *i.e.* Geometric edges or boundaries are expected where there are depth discontinuities in the depth map. Therefore, building on this idea, we utilize the edge detection operators based on the GT depth map to generate a pseudo GT edge map to calculate the loss term (BED-Loss) of Part 1 in the overall loss function. Specifically, we use the Sobel operator and Laplacian operator separately on the GT depth map to extract the first-order depth variations map and the second-order depth variations map. We then intersect these two depth variations maps and use our labeled building edge mask $\mathcal{M}$ (created using the labelme tool (wkentaro, 2018)) on the intersection to mask out non-linear features of buildings to obtain the pseudo GT edge map. The BED-Loss is defined as the mean squared error (MSE: L2 distance) between the predicted edge map and the pseudo GT edge map. By using this loss function, we can ensure that our model is accurately detecting building boundaries and producing results that are consistent with nearly real-world building edge maps. The formulation of BED-Loss is defined as follows:

$$L_{bed} = \frac{1}{N} \sum_{p \in p_{\text{valid}}} \mathcal{L}_2(E(p), \mathcal{M}(\varsigma(\Delta\hat{D}(p), \xi) \cap \phi(\Delta\hat{D}(p), \tau))), \tag{7}$$

where ς is the function that takes Sobel of the depth and threshold value ξ to return the first-order depth variations map where the Sobel response of the depth map is higher than the ξ, ϕ is the function that takes Laplacian of the depth and threshold value τ to return the second-order depth variations map where the Laplacian response of the depth map is higher than the ϕ. $\cap$ denotes the intersection operation (implemented by torch.logical_and function), N denotes the number of pixels, p_{valid} denotes the valid point set of the ground truth depth map, $\mathcal{M}$ is the labeled building edge mask. In the experimental phase, we set ξ to 4 and τ to 2 (Ibrahimli et al., 2023).

Building Surface-Depth Loss (BSD-Loss): BSD-Loss is adopted to measure the difference between the GT depth map and the predicted depth map (*e.g.* D_c, D_r) to constrain the depth estimation (Part 2). Following the previous methods (Liu and Ji, 2020; Li et al., 2023a), we also utilize the similar identical loss term to compute the average absolute value error (MAE: L1 distance) between the predicted depth map and GT depth map, called as the BSD-Loss. The BSD-Loss is defined as follows:

$$L_{bsd} = \frac{1}{N} \sum_{p \in p_{\text{valid}}} \mathcal{L}_1(\hat{D}(p), D_c(p)) + \mathcal{L}_1(\hat{D}(p), D_r(p)). \tag{8}$$

Overall loss: By using the weighted sum of the aforementioned loss terms, we create a comprehensive training criterion for our network. This approach enables us to optimize the network parameters through backpropagation. As a result, our network can learn to produce accurate and robust depth maps by minimizing the overall loss, which is a crucial factor in achieving high performance in depth estimation tasks.

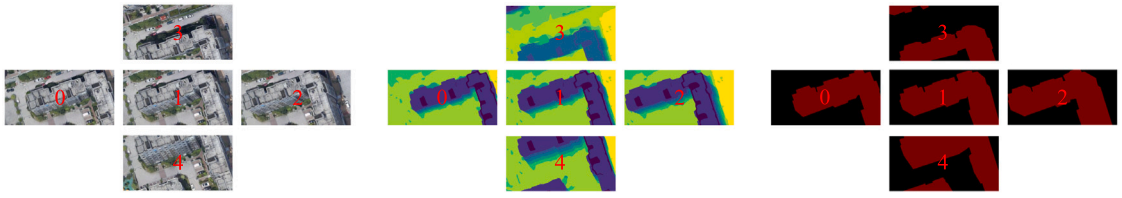
$$L_{overall} = \lambda_1 L_{bed} + \lambda_2 L_{bsd}, \tag{9}$$

where $\lambda_1 = 0.5, \lambda_2 = 0.5$ are hyper-parameters empirically set based on our experiments on the validation set.

4. Aerial building MVS dataset

This section introduces the aerial building image benckmark called the Aerial Building MVS Dataset for large-scale aerial building image depth estimation to verify the effectiveness of our proposed EG-MVSNet. It contains a set of high-resolution images of various building structures, captured from different viewpoints and different terrains. The dataset is composed of five-view images, each of which has been manually selected from the WHU dataset (Liu and Ji, 2020) and further processed using post-annotation masks (labeled building edge mask $\mathcal{M}$) to remove any unwanted objects from the images. The images in the Aerial Building MVS Dataset are aligned with the settings used in the construction of the WHU dataset (Liu and Ji, 2020), allowing for direct comparison of results between the two datasets. The benchmark is expected to provide a valuable resource for researchers working in the field of aerial building image analysis and depth estimation, and we hope that it will lead to further improvements in this important area of research.

Fig. 6. Visualization of the data set coverage region. Following the configuration of WHU Dataset (Liu and Ji, 2020), our Aerial Building MVS Dataset is mainly divided into five regions.



(a) Five-view image unit (b) Five-view depth map unit (c) Five-view labeled building edge mask unit

Fig. 7. A five-view data unit with size of 768 × 384 pixels. A five-view data unit takes the Image with ID 1 as the reference image, the images with ID 0 and 2 in the heading direction and the images with ID 3 and 4 in the side strips as the search images. From left to right: (a) five-view image unit, (b) five-view depth map unit and (c) five-view labeled building edge mask unit.

4.1. Study area and data source

We manually selected urban areas from the WHU dataset (Liu and Ji, 2020) and divided the dataset into five specific areas according to the partition scheme of WHU dataset, as shown in Fig. 6. We can observe that the main areas of our dataset are dominated by urban, with the city mainly containing dense and tall buildings, sparse factories. Therefore, this dataset is mainly used for large-scale multi-view stereo aerial building image depth estimation. Compared to the WHU dataset (Liu and Ji, 2020) and LuoJia-MVS dataset (Li et al., 2023a), our dataset is focus on the vertical domain of aerial building depth estimation, *i.e.* our dataset is designed for a specific task. Moreover, our data source comes from the WHU dataset, and we also manually annotated additional labeled building edge masks for our specific task.

4.2. Dataset construction

The Aerial Building MVS Dataset adopts the five-view data unit format, similar to the WHU dataset (Liu and Ji, 2020) and LuoJia-MVS dataset (Li et al., 2023a). Fig. 7 illustrates the five-view data unit, and the dataset contains a total of 3878 sets with overlapped views. Each view comprises an RGB image (768 × 384 pixels, 10 cm resolution), a pixel-wise depth map reprojected from 3D points, and a labeled building edge mask (post-annotation). Liu (Liu and Ji, 2020) captured the views in 11 strips with 90% heading overlap and 80% side overlap, and manually annotated each view to obtain the edge map that distinguishes foreground buildings from background terrain. The dataset comprises five representative sub-regions with different landscapes selected as training and testing sets, as visually shown in Fig. 6. The training and testing sets include various urban buildings with different terrains, using 3028 and 850 sets of five-view data units each for training and testing respectively. Our dataset is a subset of buildings from the WHU dataset (Liu and Ji, 2020). It mainly provides 8-bit RGB images, 16-bit depth maps in lossless PNG format, corresponding edge maps, and text files that record orientation parameters, including camera center (X_s, Y_s, Z_s), and rotational matrix R. The Aerial Building MVS Dataset is specifically designed for the vertical task, and as a subset of the WHU dataset, it introduces a new benchmark for MVS aerial building image depth estimation.

4.3. Other aerial MVS datasets

WHU dataset: The WHU dataset (Liu and Ji, 2020) is a large-scale aerial multi-view dataset that is captured from an oblique five-view camera rig mounted on an unmanned aerial vehicle (UAV). And using Smart3D software (Anon, 2011) to reconstruct a 3D digital surface model (DSM) with OSGB format Wang and Qian (2010) from a set of multi-view aerial images. It is a collection of artificial aerial images that was drawn from a region of roughly 6.7 × 2.2 km^2 in size that is home to many tall structures, few factories, forested mountains, bare terrain, and rivers. The sub-dataset for deep learning is made up of 4320 pairs of five-view images with a spatial resolution of 10 cm, each with a size of 768 × 384 pixels, and the ratio of the training set to test set is roughly 3:1.

LuoJia-MVS dataset: The LuoJia-MVS dataset (Li et al., 2023a) is also a large-scale aerial multi-view dataset that is generated from a 3-D DSM with OpenSceneGraph binary (OSGB) format mesh, which is built using a series of software tools, including Photoscan (Anon, 2010), Smart3D (Anon, 2011), and Meshmixer (Anon, 2009), from 1430 pairs of two-view aerial images. The dataset is also made up of 4320 pairs of five-view images with a spatial resolution of 10 cm, each with a size of 768 × 384 pixels, and the ratio of the training set to test set is roughly 3:1. The land-cover types of the LuoJia-MVS dataset are cultivated land, forest, urban areas, rural areas, industrial and mining areas, residential land, and unused land.

München dataset: The München dataset (Haala, 2014) exhibits dissimilar characteristics when compared to the WHU dataset, primarily

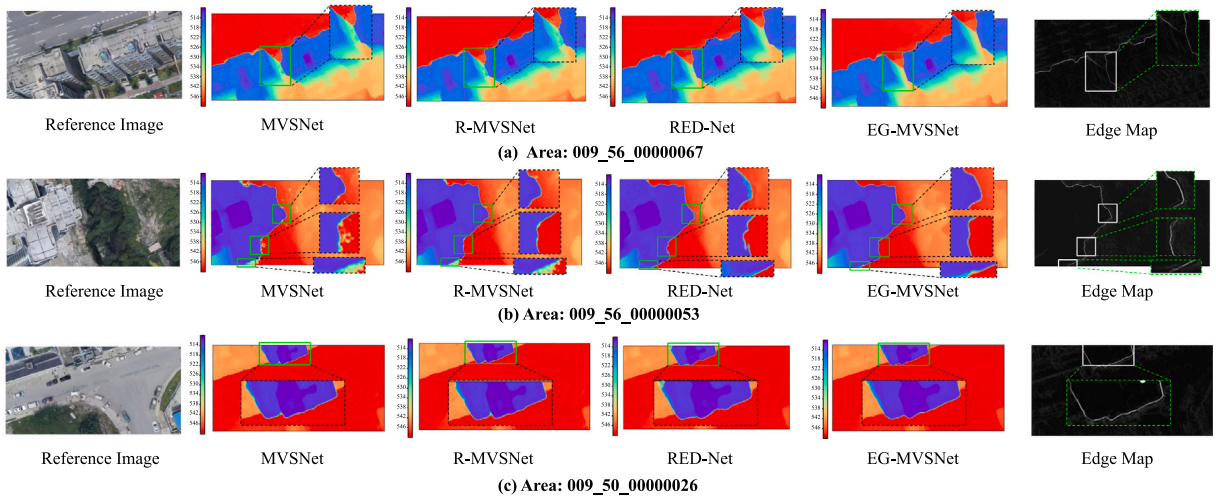


Fig. 8. Qualitative results of mainstream methods on three different areas of the Aerial Building MVS Dataset. The black dashed boxes in the figure are zoomed-in views of the local details respectively. The green dotted boxes are the corresponding edge map detail views. The scale of colorbar represents the depth value. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
The quantitative results on Aerial Building MVS Dataset. HDC-MVSNet (Li et al., 2023a) does not provide open source code.'3' and '5' represent the number of views used in inference phrase.

Views	Methods	MAE ↓	$\delta < 1.0$ m ↑	$\delta < 0.6$ m ↑	RMSE ↓	RMSE-log ↓	Abs-Rel ↓	Sq-Rel ↓	SILog ↓	Mean log10 ↓
-	COLMAP (Schonberger and Frahm, 2016)	0.2258	0.9299	0.9002	1.3908	0.0045	0.0007	0.0072	0.3393	0.0004
	ACMM (Xu and Tao, 2019)	0.1701	0.9673	0.9413	1.2093	0.0024	0.0005	0.0050	0.2477	0.0002
3	MVSNet (Yao et al., 2018)	0.2465	0.9213	0.8905	1.5901	0.0063	0.0007	0.0085	0.3530	0.0004
	R-MVSNet (Yao et al., 2019)	0.2312	0.9277	0.8980	1.4920	0.0052	0.0007	0.0078	0.3398	0.0004
	Fast-MVSNet (Yu and Gao, 2020)	0.2238	0.9309	0.9011	1.3900	0.0043	0.0007	0.0071	0.3288	0.0004
	PatchmatchNet (Wang et al., 2021)	0.2133	0.9454	0.9152	1.3111	0.0032	0.0006	0.0059	0.2998	0.0004
	RED-Net (Liu and Ji, 2020)	0.1981	0.9519	0.9203	1.2623	0.0025	0.0005	0.0052	0.2510	0.0002
	MVSNet-Cas (Gu et al., 2020)	0.1737	0.9683	0.9444	1.1008	0.0019	0.0005	0.0035	0.2174	0.0003
	EG-MVSNet (Ours)	**0.1498**	**0.9700**	**0.9498**	**0.9192**	**0.0018**	**0.0004**	**0.0026**	**0.1674**	**0.0002**
5	MVSNet (Yao et al., 2018)	0.2138	0.9408	0.9164	1.4265	0.0057	0.0006	0.0081	0.3380	0.0003
	R-MVSNet (Yao et al., 2019)	0.1976	0.9422	0.9221	1.3913	0.0049	0.0006	0.0073	0.3122	0.0003
	Fast-MVSNet (Yu and Gao, 2020)	0.1889	0.9533	0.9303	1.3023	0.0039	0.0006	0,0068	0.2877	0.0003
	PatchmatchNet (Wang et al., 2021)	0.1737	0.9617	0.9388	1.2876	0.0031	0.0006	0.0057	0.2663	0.0003
	RED-Net (Liu and Ji, 2020)	0.1670	0.9770	0.9580	1.1103	0.0022	0.0005	0.0048	0.2200	0.0002
	MVSNet-Cas (Gu et al., 2020)	0.1422	0.9809	0.9667	0.9351	0.0018	0.0004	0.0032	0.1833	0.0002
	EG-MVSNet (Ours)	**0.1252**	**0.9921**	**0.9735**	**0.7087**	**0.0014**	**0.0003**	**0.0022**	**0.1398**	**0.0001**

due to its capture location within a metropolis rather than a town setting. It comprises a collection of 15 aerial images, each boasting dimensions of 7072 × 7776 pixels, and features overlapping percentages of 80% in the heading direction and 60% in the side direction.

5. Experiment

5.1. Metrics

We employ general nine metrics to evaluate the quality of the estimated depth map. These metrics are: (1) Mean Absolute Depth Error (MAE): This is computed by averaging the $\mathcal{L}_1$ distances between the estimated and true depths. Only distances that fall within 100 depth intervals are considered to exclude extreme outliers. (2) $\delta < 1$ m: This measures the percentage of pixels whose $\mathcal{L}_1$ error was less than the 1.0 m threshold. (3) $\delta < 0.6$ m: This measures the percentage of pixels whose $\mathcal{L}_1$ error was less than the 0.6 m threshold. (4) Root Mean Square Error (RMSE): This is defined to measure the differences between the predicted and ground truth values, which are first squared, and then averaged across all pixels in the image. The square root of this average is then taken to obtain the final RMSE score. (5) RMSE-log: To calculate RMSE-log in depth map estimation, the logarithms of the predicted and ground truth depth values are first calculated. Then, the differences between these values are squared, and averaged across all pixels in the image. The square root of this average is then taken to obtain the final RMSE-log score. (6) Absolute Relative Difference (Abs-Rel): This measures the absolute difference between the predicted and ground truth depth values, normalized by the ground truth depth value. (7) Square Relative Error (Sq-Rel): This measures the squared relative difference between the predicted and ground truth depth values, normalized by the ground truth depth value. (8) SILog: The SILog error measures the relationship between points in the scene, irrespective of the absolute global scale. (9) Mean log10: This measures the difference between the logarithms of the predicted and ground truth depth values, averaged across all pixels in the image.

5.2. Implement details

Training: We have implemented our proposed EG-MVSNet using PyTorch and trained it on the Aerial Building MVS Dataset. The Aerial Building MVS Dataset is reconstructed by selecting a subset and using the Labelme (wkentaro, 2018) to label the additional masks based on the WHU-MVS dataset. In the training process, the input image resolution is set as 768 × 384 pixels, and the number of training views is set to $N = 3$. The selection of reference images and source images is the same as that of RED-Net (Liu and Ji, 2020). For training, we set the virtual hypothetical depth plane value as $D_{num} = 92$. We use the Adam optimizer with $\beta 1 = 0.9$ and $\beta_2 = 0.999$, and the initial learning

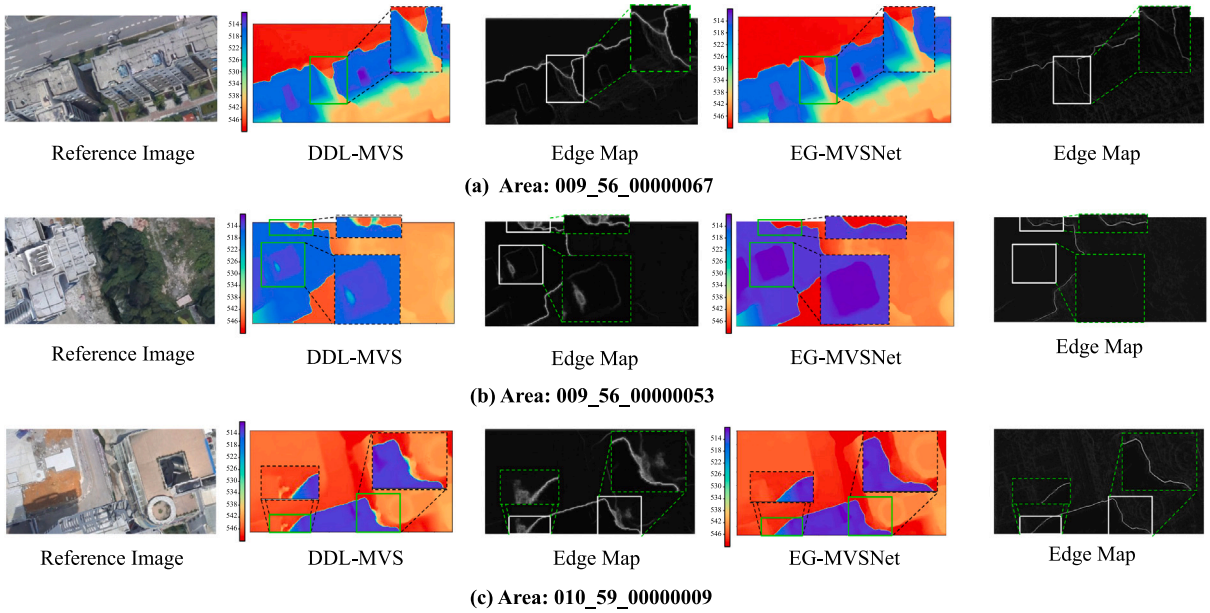


Fig. 9. Qualitative results of mainstream methods (DDL-MVS) on three different areas of the Aerial Building MVS Dataset. The green dashed boxes in the figure are zoomed-in views of the local details respectively. The withe dotted boxes are the corresponding edge map detail views. The scale of colorbar represents the depth value. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
The quantitative results on Aerial Building MVS Dataset.

Methods	MAE ↓	$\delta < 1.0$ m ↑	$\delta < 0.6$ m ↑	RMSE ↓	RMSE-log ↓	Abs-Rel ↓	Sq-Rel ↓	SILog ↓	Mean log10 ↓
PatchmatchNet (Wang et al., 2021)	0.1737	0.9617	0.9388	1.2876	0.0031	0.0006	0.0057	0.2663	0.0003
DDL-MVS (Ibrahimli et al., 2023)	0.1493	0.9804	0.9668	1.0643	0.0021	0.0004	0.0036	0.2113	0.0002
EG-MVSNet (Ours)	**0.1252**	**0.9921**	**0.9735**	**0.7087**	**0.0014**	**0.0003**	**0.0022**	**0.1398**	**0.0001**

rate is set to 0.002, with a decay weight of 0.001 applied every epoch. The model is trained for 10 epochs with a batch size of 2 and on 2× NVIDIA GTX 2080ti GPU devices. In this pipeline, only the depth map for the reference image is predicted. **Testing**: We test the best model obtained during the training process on the Aerial Building MVS test dataset, using 3 and 5 adjacent images of 768 × 384 pixels as the input respectively. The hypothetical depth plane for testing is set to $D_{num} = 92$. Finally, we adopt the aforementioned evaluation metrics to evaluate the quality of the predicted depth maps.

5.3. Benchmarking on aerial building MVS dataset

To demonstrate the effectiveness of our model, we compared our method with several traditional MVS methods and deep learning-based MVS methods. Table 1 displays the quantitative results of the Aerial Building MVS Dataset. Our observations are as follows: Firstly, our method achieved state-of-the-art overall performance when compared to other methods. Regardless of whether the input was a five-view or three-view, our method performed the best on all evaluation metrics. For instance, on the five-view results, our method improved from 0.1670 of RED-Net (Liu and Ji, 2020) to 0.1252 (25% improvement) on the MAE metric. Furthermore, our method achieved 0.9921 and 0.9735 for δ <1.0 m and δ <0.5 m respectively, for the pixel-by-pixel evaluation of the depth map. Our method produced more objective results when compared to other methods such as MVSNet (Yao et al., 2018): 0.9408 vs. R-MVSNet (Yao et al., 2019): 0.9422 vs. Our: 0.9921, RED-Net (Liu and Ji, 2020): 0.9770 vs. Our: 0.9921, MVSNet-Cas (Gu et al., 2020): 0.9809 vs. Our: 0.9921. Secondly, compared to the mainstream traditional MVS methods, e.g., COLMAP (Schonberger and Frahm, 2016) and ACMM (Xu and Tao, 2019). As detailed in Table 1, distinctly showcase the enhanced efficacy of our approach in contrast to the traditional MVS techniques. Specifically, our EG-MVSNet (5-view) exhibits a notable 44% enhancement in MAE and a 10% improvement in $\delta < 1.0$ m when compared with COLMAP. Moreover, in comparison to ACMM, our EG-MVSNet (5-view) showcases a 26% reduction in MAE and a 2% advancement in $\delta < 1.0$ m. Thirdly, the Aerial Building MVS Dataset provides reliable data support for special aerial building depth estimation and guarantees the validity of our design module. In addition, we also have conducted a qualitative comparison with MVSNet (Yao et al., 2018), R-MVSNet (Yao et al., 2019), RED-Net (Liu and Ji, 2020), and our EG-MVSNet, and present the results of our depth map visualization in Fig. 8. Although most of the estimated depth maps had similar visual quality, as seen in Fig. 8, they had difficulty in the boundary between foreground (buildings) and background (terrains) due to the depth adhesion problem, as we previously mentioned. Our method alleviates this problem effectively by incorporating edge features, resulting in better quality depth estimation (we also visualize the edge map). The quantitative results also confirm that building edge cues can improve the quality of aerial building depth map estimation.

Moreover, to validate the effectiveness of EG-MVSNet compared to other edge-based MVS methods, we also conducted comparative experiments using our Aerial Building MVS Dataset. Due to the unavailability of the EPNet (Su and Tao, 2023) source code, our comparative analysis focused on the open-source DDL-MVS framework. The results are presented in Table 2. As illustrated in Table 2, it is evident that DDL-MVS (Ibrahimli et al., 2023) demonstrates performance improvements across all evaluated metrics compared to its baseline, PatchmatchNet, upon the introduction of edge cues. However, it is noteworthy that our EG-MVSNet consistently outperforms DDL-MVS (Ibrahimli et al., 2023) in the obtained results. (MAE: 0.1252 vs. 0.1493, $\delta < 1.0$ m: 0.9921 vs. 0.9804, $\delta < 0.6$ m: 0.9735 vs. 0.9968). This superiority can be attributed to the novelty of our method, which is distinct

Table 3
The quantitative results on WHU Dataset. Some results are obtained from HDC-MVSNet (Li et al., 2023a).

Number of views	Method	MAE ↓	σ < 3-interval ↑	δ <0.6 m ↑	Comp ↑
Three-view	PatchmatchNet (Wang et al., 2021)	0.173	94.8	96.5	100%
	Fast-MVSNet (Yu and Gao, 2020)	0.184	94.1	95.5	100%
	MVSNet (Yao et al., 2018)	0.190	94.3	95.0	100%
	R-MVSNet (Yao et al., 2019)	0.183	93.5	95.3	100%
	RED-Net (Liu and Ji, 2020)	0.112	97.9	98.1	100%
	MVSNet-Cas (Gu et al., 2020)	0.111	97.6	97.7	100%
	HDC-MVSNet (Li et al., 2023a)	0.101	97.8	97.9	100%
	EG-MVSNet (Ours)	**0.097**	**98.0**	**98.2**	100%
Five-view	PatchmatchNet (Wang et al., 2021)	0.160	95.0	96.9	100%
	Fast-MVSNet (Yu and Gao, 2020)	0.157	95.6	96.1	100%
	MVSNet (Yao et al., 2018)	0.160	95.5	95.8	100%
	R-MVSNet (Yao et al., 2019)	0.173	93.8	95.4	100%
	RED-Net (Liu and Ji, 2020)	0.104	97.9	98.1	100%
	MVSNet-Cas (Gu et al., 2020)	0.095	97.8	97.8	100%
	HDC-MVSNet (Li et al., 2023a)	0.087	98.0	98.1	100%
	EG-MVSNet (Ours)	**0.081**	**98.7**	**98.5**	100%

Table 4
The quantitative results on LuoJia-MVS Dataset. Some results are obtained from HDC-MVSNet (Li et al., 2023a).

Number of views	Method	MAE ↓	σ < 3-interval ↑	δ < 0.6 m ↑
Three-view	PatchmatchNet (Wang et al., 2021)	0.252	87.2	92.7
	Fast-MVSNet (Yu and Gao, 2020)	0.194	92.0	95.7
	MVSNet (Yao et al., 2018)	0.172	92.4	96.1
	R-MVSNet (Yao et al., 2019)	0.177	93.5	96.0
	RED-Net (Liu and Ji, 2020)	0.109	96.9	98.2
	MVSNet-Cas (Gu et al., 2020)	0.103	97.1	98.4
	HDC-MVSNet (Li et al., 2023a)	0.089	97.8	98.7
	EG-MVSNet (Ours)	**0.087**	**97.9**	**98.9**
Five-view	PatchmatchNet (Wang et al., 2021)	0.283	84.1	90.4
	Fast-MVSNet (Yu and Gao, 2020)	0.357	74.9	84.6
	MVSNet (Yao et al., 2018)	0.270	81.8	91.2
	R-MVSNet (Yao et al., 2019)	0.259	86.7	92.3
	RED-Net (Liu and Ji, 2020)	0.156	90.5	94.9
	MVSNet-Cas (Gu et al., 2020)	0.141	95.4	97.9
	HDC-MVSNet (Li et al., 2023a)	0.121	96.6	98.3
	EG-MVSNet (Ours)	**0.115**	**96.9**	**98.4**

from generalized edge incorporation methods like DDL-MVS and EPNet in the context of addressing depth adhesion issues in building depth estimation. Furthermore, we have presented a qualitative analysis of our results in Fig. 9, in comparison with DDL-MVS (Ibrahimli et al., 2023), focusing on depth maps and edge maps. The visual assessment in Fig. 9 clearly showcases the efficacy of our method, particularly the ES-Net (as evidenced in Table 6), in the precise extraction of building edge features. Conversely, generalized methods such as DDL-MVS exhibit limitations as they employ generic feature extraction networks that inadvertently capture both building-specific features and extraneous, irrelevant features, exacerbating the depth adhesion issue.

5.4. Benchmarking on other aerial MVS datasets

To demonstrate the generalization performance of our method, we also validate our best model of Aerial Building MVS dataset without finetuning on other aerial image depth estimation datasets, such as WHU dataset (Liu and Ji, 2020) and LuoJia-MVS Dataset (Li et al., 2023a) Additionally, due to the lack of the open-source code of HDC-MVSNet (Li et al., 2023a), some statistics in Tables 3 and 4 are from HDC-MVSNet (Li et al., 2023a). In order to maintain fairness in comparison, we evaluated the results based on the configuration mentioned in HDC-MVSNet, which is different from the configuration mentioned in Implementation Details. Specifically, we employed the Adam optimizer with hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The dimensions of the input images were configured to 768×384 pixels. The training process was conducted for a total of 30 epochs, initialized with a learning rate of 0.001. Subsequently, the learning rate was reduced by a factor of 2 every two epochs beyond the initial ten epochs. Therefore, the results in comparison on WHU dataset are slightly higher than the results on the Aerial Building MVS Dataset (with a relatively small data gap). Furthermore, to further exhibit the generalization capacity of EG-MVSNet, we subjected it to evaluation using the München dataset (Haala, 2014), a real-world aerial dataset. Distinguished from the WHU dataset (Liu and Ji, 2020), the München dataset was captured in a metropolitan context as opposed to an urban setting. And some statistics in Table 5 are from RED-Net (Liu and Ji, 2020). Meanwhile our experimental setup aligns closely with that of RED-Net (Liu and Ji, 2020). For the München dataset (Haala, 2014), we operated with an input view number of N=3 and a depth sampling resolution of 0.1 m.

Evaluation on WHU dataset: As shown in Table 3, we test our model on WHU dataset (Liu and Ji, 2020) test set to further demonstrate the generalizability and flexibility of our EG-MVSNet. We can observe that our method achieves the highest scores in all metrics. In comparison to other methods, such as MVSNet (Yao et al., 2018), R-MVSNet (Yao et al., 2019), PatchmatchNet (Wang et al., 2021), which are not specifically designed for aerial images, our approach improves the quality of depth map estimation by introducing building edge feature information that helps the model distinguish the foreground (buildings) and the background (terrains). Table 3 demonstrates that our method achieves the state-of-the-art level in various metrics compared to other methods. For instance, compared to the ordinary MVSNet (five-view), our approach can improve the MAE from 0.160 of PatchMatch (Wang et al., 2021) to 0.081 (49% performance improvement), 0.160 of MVSNet (Yao et al., 2018) to 0.081 (49% performance improvement), and 0.173 of R-MVSNet (Yao et al., 2019) to 0.081 (53% performance improvement). Additionally, for Fast-MVSNet (Yu and Gao, 2020), our method can improve the σ < 3-interval metric from

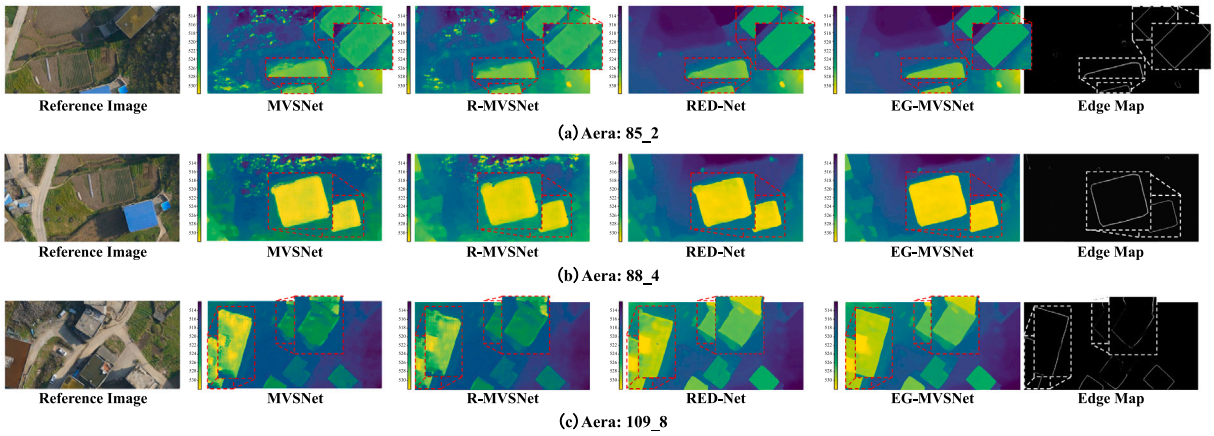


Fig. 10. Qualitative results of mainstream methods on three different areas of the LuoJia-MVS dataset. The red dashed boxes in the figure are zoomed-in views of the local details respectively. The withe dotted boxes are the corresponding edge map detail views. The scale of colorbar represents the depth value. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Quantitative evaluation on the München aerial image set (Haala, 2014) with different MVS methods. The deep learning based methods were trained on the WHU or the DTU training set. Notably, some results are obtain from the RED-Net (Liu and Ji, 2020).

Methods	Train set	MAE (m) ↓	< 3-interval (%) ↑	< 0.6m (%)
COLMAP (Schonberger and Frahm, 2016)	/	0.5860	73.36	81.95
SURE (Rothermel et al., 2012)	/	0.5138	73.71	85.70
MVSNet (Yao et al., 2018)	DTU	1.1696	43.19	61.26
	WHU-3	0.6169	69.33	81.36
	WHU-5	0.5882	70.43	83.46
R-MVSNet (Yao et al., 2019)	DTU	0.7809	43.22	70.26
	WHU-3	0.6228	74.33	83.35
	WHU-5	0.6426	74.08	83.68
RED-Net (Liu and Ji, 2020)	DTU	0.6867	63.04	78.89
	WHU-3	0.5063	80.67	86.98
	WHU-5	0.5283	80.40	86.69
EG-MVSNet (Ours)	DTU	0.6675	68.11	82.41
	WHU-3	0.4951	82.16	88.61
	WHU-5	**0.4827**	**82.56**	**89.03**

95.6 to 98.7. Furthermore, compared with other aerial image depth estimation methods such as RED-Net (Liu and Ji, 2020) and HDC-MVSNet (Li et al., 2023a), our method enhances the MAE metric from 0.104 and 0.087 to 0.081, respectively. Moreover, it is worth noting that the Aerial Building MVS Dataset is a subset of the WHU dataset (Liu and Ji, 2020), resulting in a negligible data gap. Therefore, benefiting from the mechanism of incorporating the building information to alleviate the depth adhesion, our method is capable of achieving excellent results when evaluated directly on the WHU Dataset.

Evaluation on LuoJia-MVS dataset: We conduct further experiments on other LuoJia-MVS dataset (Li et al., 2023a) to thoroughly validate the performance of EG-MVSNet in the more complex aerial environment. As shown in Table 4, we can observe that our model attains the state-of-the-art level (highest level in all metrics, MAE: 0.087, σ < 3-interval: 97.9, δ < 0.6 m: 98.9) by introducing edge feature information. Specifically, in comparison to the ordinary MVSNet (three-view) approaches, our proposed method demonstrates significant improvements in performance. The mean absolute error (MAE) is enhanced from 0.252 of PatchMatch (Wang et al., 2021) to 0.087, corresponding to a performance improvement of 65%. Similarly, the MAE is improved from 0.194 of Fast-MVSNet (Yu and Gao, 2020) to 0.087 (55% improvement), and from 0.172 of MVSNet (Yao et al., 2018) to 0.087 (49% improvement). Additionally, when compared to R-MVSNet (Yao et al., 2019), our method achieves a notable enhancement in the σ < 3-interval metric, increasing from 93.5 to 97.9. Furthermore, compared with other aerial image depth estimation methods such as RED-Net (Liu and Ji, 2020) and HDC-MVSNet (Li et al., 2023a), our method enhances the MAE metric from 0.104 and 0.087 to 0.081, respectively. However, the LuoJia-MVS Dataset (Li et al., 2023a) comprises more varied terrain scenes, leading to a disparity between these datasets. This variance affects the effectiveness of our module designed for building edges, causing our method to only perform marginally better than HDC-MVSNet on the LuoJia-MVS Dataset, as opposed to the superior results achieved on the other two datasets. Both of two tables demonstrate that our proposed method has excellent generalization performance on other aerial image depth estimation datasets. Furthermore, we have also presented the depth map visualizations as depicted in Fig. 10. While the majority of the inferred depth maps exhibit akin visual quality, they encounter challenges in accurately demarcating the boundary between foreground (buildings) and background (terrains), a predicament attributed to the inherent depth adhesion issue as discussed earlier. Our method adeptly mitigates this challenge through the integration of edge features, which engenders more precise depth estimation. This assertion is fortified by the quantitative results, affirming that the incorporation of building edge cues indeed augments the fidelity of aerial building depth map estimation.

Evaluation on München dataset: The quantitative results are presented in Table 5, offering valuable insights. Firstly, EG-MVSNet, pretrained on the WHU-5 dataset, exhibited superior performance across all evaluated metrics. Impressively, EG-MVSNet outperformed the prior state-of-the-art approach (RED-Net) by a substantial margin, specifically with an improvement of at least 2.7% in the 3-interval and

Table 6
Ablation study on the Aerial Building MVS Dataset, which demonstrates the effectiveness of different modules of our method. We remove each module from EG-MVSNet separately.

Methods	MAE ↓	$\delta < 1.0$ m ↑	$\delta < 0.6$ m ↑	RMSE ↓	RMSE-log ↓	Abs-Rel ↓	Sq-Rel ↓	SILog ↓	Mean log10 ↓
EG-MVSNet	**0.1252**	**0.9921**	**0.9735**	**0.7087**	**0.0014**	**0.0003**	**0.0022**	**0.1398**	**0.0001**
- ESNet	0.1324	0.9854	0.9719	0.7782	0.0015	0.0004	0.0025	0.1538	0.0003
- EPB	0.1351	0.9858	0.9632	0.7871	0.0017	0.0005	0.0031	0.1617	0.0004
- IAFM	0.1319	0.9862	0.9717	0.7346	0.0015	0.0004	0.0024	0.1442	0.0003
- EDRM	0.1297	0.9892	0.9721	0.7132	0.0015	0.0003	0.0023	0.1417	0.0002

Table 7
Ablation study on the Aerial Building MVS Dataset, which demonstrates the effectiveness of DSK module. 'SC' means replacing the DSK with the standard convolution.

Methods	MAE ↓	$\delta < 1.0$ m ↑	$\delta < 0.6$ m ↑	RMSE ↓	RMSE-log ↓	Abs-Rel ↓	Sq-Rel ↓	SILog ↓	Mean log10 ↓
+ DSK (Ours)	**0.1252**	**0.9921**	**0.9735**	**0.7087**	**0.0014**	**0.0003**	**0.0022**	**0.1398**	**0.0001**
+ SC	0.1298	0.9868	0.9722	0.7643	0.0015	0.0004	0.0024	0.1507	0.0003

Table 8
Ablation study on the Aerial Building MVS Dataset, which demonstrates the effectiveness of IAFM module. 'ISEM' means incorporating the edge features and surface features via the SE block (Hu et al., 2018).

Methods	MAE ↓	$\delta < 1.0$ m ↑	$\delta < 0.6$ m ↑	RMSE ↓	RMSE-log ↓	Abs-Rel ↓	Sq-Rel ↓	SILog ↓	Mean log10 ↓
+ IAFM (Ours)	**0.1252**	**0.9921**	**0.9735**	**0.7087**	**0.0014**	**0.0003**	**0.0022**	**0.1398**	**0.0001**
+ ISEM	0.1295	0.9890	0.9723	0.7298	0.0015	0.0004	0.0023	0.1428	0.0002

Table 9
Ablation study on the Aerial Building MVS Dataset, which demonstrates the effectiveness of EDRM module. 'DMR' represents the Depth Map Refinement used in MVSNet.

Methods	MAE ↓	$\delta < 1.0$ m ↑	$\delta < 0.6$ m ↑	RMSE ↓	RMSE-log ↓	Abs-Rel ↓	Sq-Rel ↓	SILog ↓	Mean log10 ↓
+ EDRM (Ours)	**0.1252**	**0.9921**	**0.9735**	**0.7087**	**0.0014**	**0.0003**	**0.0022**	**0.1398**	**0.0001**
+ DMR	0.1258	0.9913	0.9727	0.7099	0.0015	0.0003	0.0023	0.1402	0.0001

2.6% in the < 0.6 m. This notable enhancement can be attributed to the distinctive architectural emphasis of EG-MVSNet on capturing building edge features. The adeptness of EG-MVSNet in handling urban building scenes, a characteristic prominently featured in the München dataset (Haala, 2014), due to it mainly consist of many urban building scenes, has manifested in its remarkable generalization prowess. Consequently, EG-MVSNet demonstrated the most favorable generalization performance, characterized by an MAE of 0.4827, a < 3-interval of 82.56, and a < 0.6 m of 89.03. Secondly, in contrast to RED-Net's propensity for better performance on WHU-3 compared to WHU-5, EG-MVSNet exhibited a different trend. Notably, EG-MVSNet showcased better performance on WHU-5 relative to WHU-3. This behavior can be attributed to the utilization of more images for training in the case of WHU-5. This augmentation enables the ESNet of EG-MVSNet to be more effectively trained, consequently leading to its superior performance on the München dataset (Haala, 2014), which prominently buildings.

5.5. Ablation study

In this section, we have conducted an ablation study to understand and analyze the contributions of the modules of our architecture. The results are shown in Table 6.

Effectiveness of ESNet: We replaced the ESNet with SFENet in our EG-MVSNet to extract the features for the subsequent IAFM, EDB, and EDRM. The experimental results, shown in Table 6, indicate that while SFENet extracts some useful information for the subsequent modules, its ability to capture building edge information is insufficient. This leads to the incorporation of surface information as noise, affecting the quality of depth map estimation. In particular, we observed that the introduction of ESNet leads to a more accurate and robust feature extraction process. This is reflected in the improved performance metrics, such as MAE, $\sigma < 1.0$ m, $\delta < 0.6$ m, and so on, as shown in Table 6 Row 1 and Row 2. The quantitative results effectively demonstrate the validity of our ESNet.

Effectiveness of DSK: Our proposed an Edge-Sensitive Network (ESNet) is based on the differentiable Dynamic Sobel Kernels (DSK) to capture the rigid features of building. Therefore, to further validate the effectiveness of our DSK module, we also have conducted an ablation experiment with DSK or without DSK. As shown in Table 7, it is evident that upon substituting our DSK module with standard convolution, there is a degradation across all metric aspects (e.g., MAE: 0.1252 → 0.1298, $\delta < 1.0$ m: 0.9921 → 0.9868). This outcome can be attributed to the DSK module's adeptness at encoding linear features effectively. In contrast, standard convolution, while capable of capturing linear features, also encompasses a broader range of surface texture features. This multiplicity of features could potentially impede the semantic representation of edge information.

Effectiveness of EPB: We directly remove the entire EPB and then directly use the building edge features extracted by ESNet for subsequent stages. EPB serves as one of the core modules of our network, which is able to ensure that our ESNet extracts accurate and reliable building edge features for subsequent stages. While ESNet is effective in capturing edge features, the lack of BED-Loss constraint in EPB results in the extraction of many non-edge linear features, such as surface textures on the object surface. Thus, upon removing the EPB, MAE decreases from 0.1252 to 0.1351, $\sigma < 1.0$ m from 0.9921 to 0.9858, $\delta < 0.6$ m from 0.9735 to 0.9632, as shown in Table 6 Row 1 and Row 3. The quantitative results effectively demonstrate the validity of our EPB.

Effectiveness of IAFM: We replaced our proposed IAFM with a method that involved summing the edge feature volume and the standard cost volume. The experimental results are shown in Table 6 Row 1 and Row 4. Although the summation method had some effect on the results, the final estimated depth map still had a depth sticking problem due to the lack of enhancement of the cost information by the edge feature information. This, in turn, affected the overall quality of the depth estimation. To further exhibit the superiority of our IAFM, we have adopted an ISEM module (incorporate the edge features and surface features via the SE block (Hu et al., 2018)) to replace the IAFM module to conduct an comparative experiments. From Table 8 we can observe that the replacement of the edge feature volume and standard cost volume with the incorporation of surface features and edge features obtains worse results in experiments, e.g., MAE: 0.1252 vs. 0.1295, $\delta <$

Table 10
Performance comparisons on the Aerial Building MVS evaluation dataset. "Run-time" represents the inference time.

Methods	Input Size	View	Num Depth	MAE	$\delta < 1.0$ m	$\delta < 0.6$ m	GPU Mem (MB)	Run-time (s)	Output Size
MVSNet (Yao et al., 2018)	768 × 384	3	94	0.2465	0.9213	0.8905	2401 MB	0.2510 s	192 × 96
R-MVSNet (Yao et al., 2019)	768 × 384	3	94	0.2312	0.9277	0.8980	1993 MB	1.3244 s	192 × 96
RED-Net (Liu and Ji, 2020)	768 × 384	3	94	0.1981	0.9519	0.9203	3199 MB	0.8139 s	768 × 384
EG-MVSNet (Ours)	768 × 384	3	94	**0.1498**	**0.9700**	**0.9498**	4567 MB	1.2325 s	768 × 384

1.0 m: 0.9921 vs. 0.9890, $\delta < 0.6$ m: 0.9735 vs. 0.9723. Our findings illustrate the superiority of 3D volume incorporation over 2D feature integration. This enhancement is attributed to augmented information dimensionality (3D volume-level), compared to 2D feature-level fusion, which can causes some semantic information loss. After the information upgrading, can add additional depth information, more conducive to the information fusion based on the attention mechanism, and therefore will bring some performance improvement. The quantitative results effectively demonstrate the validity of our IAFM.

Effectiveness of EDRM: We removed EDRM directly from EG-MVSNet, and the resulting metrics are presented in Table 6. A comparison between Row 1 and Row 5 reveals that the incorporation of edge features to refine the coarse depth map significantly improves the model's metrics, including MAE: 0.1297 to 0.1252, $\sigma < 1.0$ m: 0.9892 to 0.9921, $\delta < 0.6$ m: 0.9721 to 0.9735. To further demonstrate the superiority of our EDRM, we have conducted an additional ablation study to compare the performance between our EDRM with the Depth Map Refinement (DMR) used in MVSNet. From Table 9, we can observe that comparing the EDRM and DMR, our EDRM achieves better performance in all evaluation metrics (MAE: 0.1252 vs. 0.1258, $\delta < 1.0$ m: 0.9921 vs. 0.9913, $\delta < 0.6$ m: 0.9735 vs. 0.9727, RMSE: 0.7087 vs. 0.7099) on the Aerial Building MVS Dataset. The key distinction between our EDRM and DMR, lies in the comprehensive integration of edge features in our EDRM module to address the depth adhesion issue. This feature allows our EDRM to refine the depth map with enhanced efficacy compared to DMR, which solely relies on surface features. As a result of this edge-aware refinement, our EDRM manifests improvements across all metrics when contrasted with DMR. The quantitative results effectively demonstrate the validity of our EDRM.

5.6. GPU and runtime

In this section, we have conducted an efficiency experiment to compare the runtime and memory with the other methods. As shown in Table 10, we have two observations: Firstly, we can observe that MVSNet demonstrated the shortest inference time when compared to all the methods included in the study. This efficiency can be attributed to the strategic utilization of 4x downsampling features implemented throughout the entire pipeline. These downsampling features significantly reduce the computational overhead. However, our method still outperforms than other methods in all evaluation metrics (MAE:0.1498, $\delta < 1.0$ m:0.9700, $\delta < 0.6$ m:0.9498) while delivering larger depth maps. This advantageous performance can be attributed to the effectiveness of incorporating the edge cues into MVS pipeline and the efficiency of our Edge Prediction Branch (EPB), which employs a simple yet effective 2D UNet architecture. Importantly, this design choice does not introduce significant computational overhead compared to our baseline, e.g., 0.8139 s → 1.2325 s. Secondly, both EG-MVSNet and RED-Net exhibit higher memory requirements compared to MVSNet and R-MVSNet. This observation can be attributed to the feature extraction network adopted by MVSNet and R-MVSNet, yielding 4x downsampling features used through the entire pipeline. This downsampling strategy effectively curtails the memory requisites. In contrast, our EG-MVSNet and RED-Net maintain output sizes of depth map equivalent to the input size of images, consequently using an enlarged scale of tensors throughout the pipeline, which causes the larger the memory requirements. Although our EG-MVSNet requires more memory than RED-Net under the same configurations, due to the edge prediction branch, which increases more memory requirements (3199MB → 4567 MB,). However, also benefit from the edge prediction branch, our EG-MVSNet outperforms RED-Net on most evaluation benchmarks by incorporating edge cues into MVS to alleviate the problem of depth adhesion, e.g., MAE:0.1498, $\delta < 1.0$ m:0.9700, $\delta < 0.6$ m:0.9498.

5.7. Visualization results

In this section, we present visualization results to validate the effectiveness of our proposed modules. As depicted on the left-part side of Fig. 11, we incorporate the features of the reference image as a mask to the image. The orange portions represent the valid features in the mask, while the blue portions indicate some invalid areas. Therefore, the left-part side of Fig. 11 demonstrates that our ESNet can effectively extract accurate and reliable edge features. Meanwhile, to confirm the reliability of the depth map regressed by our EPB, we mask the edge map to the reference image. As demonstrated on the right-part side of Fig. 11, our edge map aligns perfectly with the edges of the building, thus proving the effectiveness of our proposed BED-Loss constraint in constraining the EPB to predict a precise and reliable edge map.

5.8. Aerial building area reconstruction

Our EG-MVSNet is capable of generating full-resolution depth maps with an arbitrary number of depth hypothesis planes, making it highly beneficial for reconstructing high-resolution urban clusters from large-scale multi-view aerial images. The model can process five-view images measuring 768 × 384 pixels on a 2080ti GPU, and it takes only 0.3 s to infer a depth map with 92 depth hypothesis planes. As shown in Fig. 12, we present the reconstruction results for four areas, each covering approximately 1.8 × 0.85 km^2, and it took around 6.5 min to reconstruct a single scene. Overall, the capabilities of our EG-MVSNet hold great promise for advancing the field of aerial reconstruction and providing new insights into the urban landscape.

6. Limitations and 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, although our model achieves state-of-the-art performance than most of mainstream methods on the three benchmarks, it has several major limitations: (1) Currently, the layout of buildings in aerial multi-view stereo (MVS) benchmark images poses limitations on the effective extraction of very close building edges, despite incorporating edge-aware building boundary extraction modules. This limitation may impact the fusion of building edge information into MVS and consequently affect the accuracy of building depth estimation. To address this issue, the integration of a new branch that utilizes extracted building edge features to segment the building surfaces and distinguish between different buildings can effectively alleviate the aforementioned limitation. (2) During the construction of large-scale aerial multi-view stereo (MVS) self-constructed benchmarks, the usage of COLMAP to derive camera parameters based on real outdoor scenes may introduce potential inaccuracies. MVSNet utilizes differentiable homography warping based on these camera parameters to construct the cost volume. However, the presence of inaccurate camera parameters can still impact cost matching, leading to compromised quality

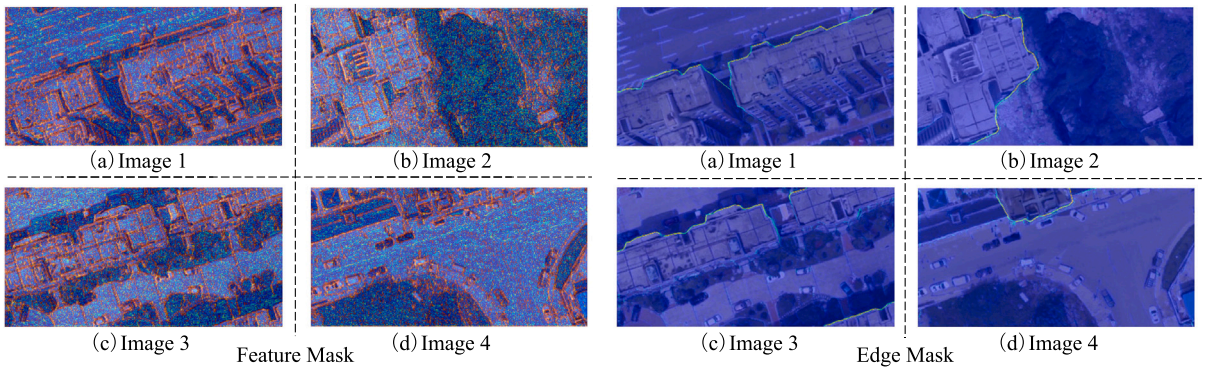


Fig. 11. Visualization results of Feature Mask and Edge Mask to validate the effectiveness of the proposed modules. In the left part, the orange portions represent the valid features in the mask, while the blue portions indicate some invalid areas. The results of aligning the Edge map on the reference image are exhibited in the right part.

Fig. 12. The point cloud reconstructions of for areas. Orange and green are different views of different areas respectively.

in depth map estimation. To mitigate this issue, the integration of a module capable of optimizing camera parameters for inaccurate views holds the potential to yield more precise camera parameters. (3) In the case of scenes characterized by intricate edges and a profusion of linear elements, we acknowledge that exclusive dependence on our DSK module might prove insufficient to comprehensively encompass exceedingly intricate linear features. Specifically, certain complex linear attributes, such as the linear edges of slanted surfaces, may not be fully extracted. To address above issue, we envisage a prospective avenue of research that entails the integration of a module capable of extracting linear features in all conceivable directions. This strategic augmentation holds the promise of effectively resolving the aforementioned issue. (4) In the context of edge-aware MVS depth inference, the precision of the edge map plays a pivotal role in determining the overall quality of the depth map. Currently, conventional edge-aware MVS methods lack the evaluation metrics for assessing the edge map's quality. Therefore, it is advisable for future research endeavors to consider generating the ground truth edge maps by tools like Labelme. This approach would empower future studies to conduct thorough evaluations of the edge map's quality, ultimately ensuring enhanced performance in depth estimation tasks.

7. Conclusion

In this paper, we propose an edge information guided depth inference network for large-scale aerial building multi-views stereo, called **EG-MVSNet**, which alleviates the problem of depth adhesion by introducing the edge feature information, thus further improving the accuracy of the estimated depth maps. The network comprises an Edge-Sensitive Network for building edge feature extraction, an UNet-like Edge Prediction Branch for building edge map regression, and a Building Edge-Depth Loss for extracting valid building edge information. An Inter-volume Adaptive Fusion Module incorporates the edge feature volume into the standard cost volume, and an Edge Depth Refinement Module further refines the depth map using the edge features. In order to validate the effectiveness of our model, we reconstruct a aerial building benchmark referred to as the Aerial Building MVS Dataset based on the WHU dataset. Through experimentation on established datasets such as WHU dataset (Liu and Ji, 2020) and LuoJia-MVS dataset (Li et al., 2023a), as well as our Aerial Building MVS Dataset, we demonstrate that our proposed method outperforms the state-of-the-art MVSNet approaches. Specifically, our method achieves the top performance in terms of SOTA MAE score, $\sigma <$ 3-interval score, and $\delta <$ 0.6 m score on both the WHU and LuoJia-MVS datasets. Moreover, our method produces the best results on our Aerial Building MVS Dataset for the specialized task of building depth estimation, as evaluated by various metrics such as MAE, $\delta <$ 1.0 m, and $\delta <$ 0.6 m. Additionally, our method exhibits superior visualization capabilities for aerial building scenes, particularly in the edge regions, on our Aerial Building MVS Dataset. In the future work, we intend to investigate the integration of EG-MVSNet with straightforward instance segmentation branches, such

as Mask R-CNN (He et al., 2017), in order to enable precise differentiation of various buildings and achieve high-precision aerial building depth estimation. Furthermore, we plan to enhance the robustness and accuracy of our method by incorporating dense bundle adjustment, which optimizes camera poses and depth maps simultaneously. This addition is essential as inaccurate camera parameters can significantly impact the performance of EG-MVSNet.

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.

References

Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B., 2016. Large-scale data for multiple-view stereopsis. Int. J. Comput. Vis. 120, 153–168.

Anon, 2009. Meshmixer, Available https://www.autodesk.com/.

Anon, 2010. Photoscan, Available https://www.agisoft.com/.

Anon, 2011. Smart 3D, Available: https://www.bentley.com/en/products/brands/contextcapture.

Anon, 2015. ContextCapture, Available: https://Www.Bentley.Com/En/Products/Brands/Contextcapture.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2020. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arxiv 2014. arXiv preprint arXiv:1406.1078.

Duan, P., Kang, X., Li, S., Ghamisi, P., Benediktsson, J.A., 2019. Fusion of multiple edge-preserving operations for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 57 (12), 10336–10349. http://dx.doi.org/10.1109/TGRS.2019.2933588.

Fang, F., Li, J., Zeng, T., 2020. Soft-edge assisted network for single image super-resolution. IEEE Trans. Image Process. 29, 4656–4668. http://dx.doi.org/10.1109/TIP.2020.2973769.

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.

Haala, N., 2014. EuroSDR-Project Commission 2 "Benchmark on image matching". Final Report, Wien, Austria.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Ibrahimli, N., Ledoux, H., Kooij, J.F., Nan, L., 2023. DDL-MVS: Depth discontinuity learning for multi-view stereo networks. Remote Sens. 15 (12), 2970.

Li, J., Huang, X., Feng, Y., Ji, Z., Zhang, S., Wen, D., 2023a. A hierarchical deformable deep neural network and an aerial image benchmark dataset for surface multi-view stereo reconstruction. IEEE Trans. Geosci. Remote Sens.

Li, A., Jiao, L., Zhu, H., Li, L., Liu, F., 2021. Multitask semantic boundary awareness network for remote sensing image segmentation. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Li, L., Zhou, Z., Wu, S., Cao, Y., 2023b. Multi-scale edge-guided learning for 3D reconstruction. ACM Trans. Multimed. Comput. Commun. Appl. 19 (3), 1–24.

Liu, S., Ding, W., Liu, C., Liu, Y., Wang, Y., Li, H., 2018. ERN: Edge loss reinforced semantic segmentation network for remote sensing images. Remote Sens. 10 (9), 1339.

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.

Qi, X., Liu, Z., Liao, R., Torr, P.H., Urtasun, R., Jia, J., 2020. Geonet++: Iterative geometric neural network with edge-aware refinement for joint depth and surface normal estimation. IEEE Trans. Pattern Anal. Mach. Intell. 44 (2), 969–984.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp. 234–241.

Rothermel, M., Wenzel, K., Fritsch, D., Haala, N., 2012. SURE: Photogrammetric surface reconstruction from imagery. In: Proceedings LC3D Workshop, Vol. 8, No. 2. Berlin.

Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J.D., Breitkopf, U., Jung, J., 2014. Results of the ISPRS benchmark on urban object detection and 3D building reconstruction. ISPRSJ. Photogramm. Remote Sens. 93, 256–271.

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

Sobel, I., Feldman, G., et al., 1968. A 3 × 3 isotropic gradient operator for image processing. In: A Talk at the Stanford Artificial Project. pp. 271–272.

Song, X., Zhao, X., Fang, L., Hu, H., Yu, Y., 2020. Edgestereo: An effective multi-task learning network for stereo matching and edge detection. Int. J. Comput. Vis. 128, 910–930.

Su, W., Tao, W., 2023. Efficient edge-preserving multi-view stereo network for depth estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, No. 2. pp. 2348–2356.

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, R., Qian, X., 2010. OpenSceneGraph 3.0: Beginner's Guide. Packt Publishing Ltd.

Wei, Z., Ding, S., Cheng, L., Xu, W., Wang, Y., Zhang, L., 2022. Linear building pattern recognition in topographical maps combining convex polygon decomposition. Geocarto Int. 1–25.

wkentaro, 2018. labelme, https://github.com/wkentaro/labelme.

Xiang, X., Wang, Z., Lao, S., Zhang, B., 2020. Pruning multi-view stereo net for efficient 3D reconstruction. ISPRS J. Photogramm. Remote Sens. 168, 17–27.

Xie, S., Tu, Z., 2015. Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1395–1403.

Xu, W., Guangluan, X., Wang, Y., Sun, X., Lin, D., Yirong, W., 2018a. High quality remote sensing image super-resolution using deep memory connected network. In: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 8889–8892.

Xu, Q., Tao, W., 2019. Multi-scale geometric consistency guided multi-view stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5483–5492.

Xu, W., Wang, J., Wang, Y., Xu, G., Lin, D., Dai, W., Wu, Y., 2020. Where is the model looking at?–Concentrate and explain the network attention. IEEE J. Sel. Top. Sign. Proces. 14 (3), 506–516.

Xu, W., Wang, J., Wei, Z., Peng, M., Wu, Y., 2023. Deep semantic-visual alignment for zero-shot remote sensing image scene classification. ISPRS J. Photogramm. Remote Sens. 198, 140–152.

Xu, W., Xian, Y., Wang, J., Schiele, B., Akata, Z., 2022. Attribute prototype network for any-shot learning. Int. J. Comput. Vis. 130 (7), 1735–1753.

Xu, W., Xu, G., Wang, Y., Sun, X., Lin, D., Wu, Y., 2018b. Deep memory connected neural network for optical remote sensing image restoration. Remote Sens. 10 (12), 1893.

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. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV. Springer, pp. 674–689.

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. 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, Z., Feng, C., Liu, M.-Y., Ramalingam, S., 2017. Casenet: Deep category-aware semantic edge detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5964–5973.

Yu, Z., Gao, S., 2020. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1949–1958.

Yu, A., Guo, W., Liu, B., Chen, X., Wang, X., Cao, X., Jiang, B., 2021a. 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., 2021b. Automatic 3D building reconstruction from multi-view aerial images with deep learning. ISPRS J. Photogramm. Remote Sens. 171, 155–170.

Zhu, Z., Stamatopoulos, C., Fraser, C.S., 2015. Accurate and occlusion-robust multi-view stereo. ISPRS J. Photogramm. Remote Sens. 109, 47–61.