Deep-FG-DSM: Fine-Grained DSM for High-resolution UAV Imagery Based on Deep Implicit Occupancy Networks

Yongmao Hou, Haibo Zhang, Xin Wang*, Zongqian Zhan

School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China
2020302142004@whu.edu.cn, 2120450368@qq.com, (xwang, zqzhan)@sgg.whu.edu.cn

Keywords: Unmanned Aerial Vehicle (UAV) Imagery, Digital Surface Model (DSM), Occupancy Networks, High resolution

Abstract

Nowadays, DSM (Digital Surface Model) is one of most important products that has been widely applied in digital city or smart city. Over the last decades, the common way is to follow the conventional photogrammetric pipeline for generating DSM, which is often not efficient and yields noise with small geometric details lost. Inspired by the development of deep implicit occupancy network, in this paper, we presented a learning-based method for obtaining fine-grained DSM, i.e., fine Deep-FG-DSM. In particular, high-resolution UAV imagery together with the corresponding original point cloud are employed to improve DSMs preserving higher details, two heads that embed the features of images and 3D points are deployed, and a MLP (Muti-layer Perceptron) is appended to decode these embedding into continuous occupancy probabilities for predicting the existence (or not) of surface point. Our experimental results demonstrate the robustness of our model against both sparse and noisy point clouds. While generating DSMs, it retains high-frequency details from high-resolution UAV images while maintaining relatively high accuracy. For point cloud obtained after simplification with average sampling resolution of d=5m, the MAE (Mean Absolute Residual Error) is 2.15m.

1. Introduction

In the last few years, the product of DSM has been extensively used as fundamental geographic information data to support digital city and smart city. To generate DSM, recently, thanks to the corresponding flexibility and high efficiency, high-resolution cameras installed on UAVs (unmanned aerial vehicles) have been widely used and conventional photogrammetric procedures are adopted. However, these procedures (including image matching, image orientation, two-view stereo dense matching, point cloud fusion etc.) typically results in significant noise and fails to preserve the fine details that are often clearly presented on UAV images. Recently, learning-based implicit representation studies have gained increasing popularity, such as DeepSDF (Park et al., 2019) and IMPLICITY (Stucker et al., 2022), which, in principle, can yield infinite resolution DSM, and its practical resolution might be limited by factors such as input data resolution and the number of neurons in the network architecture. Thus, learning-based implicit representation could be a promising alternative for generating fine-gained DSM in complex and large-scale scenes.

Recently, several works have been tried to explore the possibility of implicit representation for mesh model. For example, PIFu (Saito et al., 2019) proposed pixel-aligned implicit functions that are capable to reconstruct detailed 3D models of clothed human bodies from a single image, preserving high-frequency details like clothing folds. However, this method is originally designed for modeling individual small sized objects, especially for human body. Later, learned by the idea of occupancy map, Convolutional Occupancy Networks (CON) (Peng et al., 2020) addressed this limitation by introducing inductive bias and convolutional operations. CON refined object surface modeling by predicting the occupancy probability of input points, extending from single-object to implicit 3D reconstruction in large-scale indoor scenes. The latest work, IMPLICITY (Stucker et al., 2022), combined the network architectures of CON and PIFu for producing large area DSM from satellite imagery. Although IMPLICITY preserves local details of buildings, it focuses solely on the 3D point features, neglecting the local characteristics among neighboring points, which may lead to deformations in building shapes and rough edges for very sparse point cloud, deviating from their authentic forms.

Based the aforementioned methods, this paper proposes Deep-FG-DSM for high-resolution UAV imagery, a novel method for fine-grained DSM generation which is robust to sparse and noisy point cloud. Fig. 1 shows the general workflow of our method. Specifically, in the embedding phase, two branches are employed to encode the information of images and 3D points. In the 3D point encoder, the backbone of PointNet++ is adopted, the local neighboring points are investigated to improve the feature representation of the input point, which compromises between local details and global accuracy. Additionally, more informative RGB Ortho-images are integrated via an image encoder to guide occupancy probability predictions, addressing challenges posed by detailed structures in buildings.

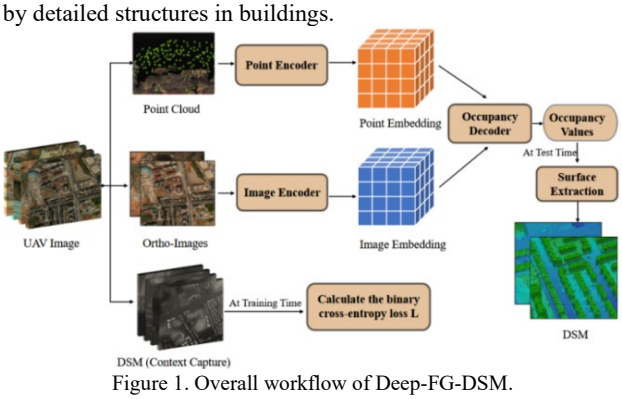


Figure 1. Overall workflow of Deep-FG-DSM.

The main contributions of this work are threefold:

1. We propose learning-based architecture for generating fine-grained DSMs using high-resolution UAV imagery, which is robust to sparse and noisy point cloud.

* Corresponding Author

2. In the point encoder, PointNet++ is deployed for extracting point cloud features.
3. This work establishes benchmark with large-scale high-resolution UAV images that is tailored for learning implicit DSM, namely, UAVIIR[1] (Unmanned Aerial Vehicle Imagery Implicity Reconstruction).

The rest of this paper is organized as follows: Section 2 discusses related work. The details of the proposed method are illustrated in Section 3. The performance of our work and the experimental results are reported in Section 4. Finally, conclusions and an outlook are drawn in Section 5.

2. Related Work

In this section, we review some relevant methods of generating DSM. In general, these works can be mainly divided into traditional methods and deep learning-based methods, in which learning-based 3D methods can be further categorized into implicit and explicit methods based on the employed 3D representation. Comparing to explicit method, implicit deep learning-based methods are demonstrated to be able to achieve higher accuracy and resolution. However, currently, most implicit methods are limited on reconstructing small-scale scenes or single objects with simple shapes.

Traditional methods. The traditional DSM generation using photogrammetric pipeline typically takes the results of stereo dense matching as input, which can be classified into three categories: local matching, global matching, and semi-global matching (SGM). Stereo dense matching aims to find a dense (even pixel by pixel) set of correspondences from image pairs, followed by computing the disparity for each pixel, and ultimately converting disparity into depth values. A seminal work is the multi-stage optimization framework proposed by (Scharstein and Szeliski, 2002), it consists of four crucial stages: matching cost computation, measuring pixel similarity in stereo image pairs; matching cost aggregation, aggregating local matching costs to improve robustness; disparity calculation, estimating disparities based on aggregated costs; and finally, disparity refinement, enhancing the accuracy of the estimated disparities. Local stereo matching algorithms improve computational efficiency by aggregating matching costs within a local window and determining the optimal disparity for each pixel. Some popular works of this category include Window-based Matching (Kanade et al., 1994), Block Matching (Dabov et al., 2006), and Normalized Cross-Correlation (Briechle et al., 2001). Global stereo matching algorithms consider global information within the images to ensure consistency and coherence across the entire disparity map, it transforms the problem of disparity calculation into an energy optimization task to find the best matching pixel pairs, thus obtaining the disparity values for each pixel, for example, Graph Cuts (Boykov et al., 2006), Random Walks (Shen et al., 2008). Compared to traditional local and global matching algorithms, Semi-Global Matching (SGM) integrates the advantages of both by considering both local region information and utilizing global consistency. It calculates the disparity for each pixel by aggregating in multiple directions on the pixel grid and optimizes an energy function to obtain the final disparity map. Some significant works in this direction include Semi-Global Block Matching (Gehrig et al., 2009), a variant of SGM that employs techniques such as pixel stratification and rapid cost aggregation, and Efficient Large-Scale Stereo Matching (Geiger et al., 2010), which utilizes techniques of multi-scale matching and gradient-based cost to handle large-scale stereo matching problems.

Deep Explicit Methods. The relevant deep learning-based methods that output explicit 3D representation, such as voxels, point clouds, and meshes as output representations, are referred as deep explicit methods. These works explicitly define the geometric shape and topological structure of reconstructed 3D objects, presenting scenes accurately and describing the shape, position, and orientation of objects or scenes precisely. Voxel representation is one of the earliest 3D representations applied in deep learning-based studies. Choy et al. (2016) proposed D3-R2N2, which is a recurrent neural network architecture taking sequential images as input, it is capable of outputting the reconstruction results of objects in the form of 3D voxels from arbitrary viewpoints. Similarly, 3D-GAN (Wu et al., 2016), built upon volumetric convolutional neural networks and generative adversarial networks, generates voxel representations of 3D objects from probabilistic space. An alternative common 3D representation is 3D points, such as, Fan et al. (2017) proposed the Point Set Generation Network (PSGN), which aims to generate a 3D point cloud representation from a single image. Similarly, Yang et al. (2019) proposed a method based on continuous normalizing flows called Pointflow, which is utilized for generating high-quality 3D point clouds. Another popular 3D mesh representation methods, that employ neural networks to directly regress the vertices and faces of the mesh, have been extensively studied by Gkioxari et al. (2019), Kanazawa et al. (2018), and Lin et al. (2019). Gkioxari et al. (2019) introduced a novel method named Mesh R-CNN for reconstructing the 3D mesh representation of objects from a single image. Kanazawa et al. (2018) presented a learning framework for recovering the 3D shape, camera, and texture of an object from a single image. Additionally, Lin et al. (2019) proposed a method based on video alignment for reconstructing the 3D mesh model of objects from multiple videos.

Deep Implicit Methods. Implicit methods utilize neural networks to learn and embed spatial and structural information of complex scenes, establishing mappings between spatial information and attribute information without the explicit specification of features or rules. They enable querying attribute information such as color, occupancy, distance (etc.) of any point in space via spatial coordinates. Compared to the explicit representations that require spatial discretization (e.g., based on voxels, point clouds, and meshes), implicit models can continuously represent various shapes and inherently handle complex topologies, thus capable of generating higher-resolution reconstruction results. Recently, many studies work on implicit occupancy fields (Mescheder, Lars, et al. 2019; Chen, Zhiqin et al. 2019) and distance fields (Michalkiewicz, Mateusz, et al. 2019; Park, Jeong Joon, et al. 2019), which train neural networks to infer the occupancy probability or distance value for any given 3D point, and then using Marching Cubes algorithm to extract iso-surfaces to obtain the 3D model of the object. To the best of our knowledge, a groundbreaking work in the field of deep implicit 3D reconstruction is the Pixel-aligned Implicit Function (PIFU) proposed by Saito et al. (2019), PIFU associates pixels in 2D images with corresponding 3D information of human bodies, enabling the reconstruction of detailed 3D models of clothed human bodies from single images while preserving high-frequency details such as clothing folds. Additionally, this work introduces a sampling strategy for training space implicit functions that combines uniform sampling with adaptive sampling based on surface geometry. Another popular work is

[1] More details related to UAVIIR will be found at Section 4.1.

Convolutional Occupancy Networks (CON) proposed by Peng et al. (2020). Prior to this work, most deep implicit methods were limited by the simple fully connected network architectures, which is unable to integrate local features or merge inductive biases, and could only perform 3D reconstruction on small scenes or simple objects. CON addresses these limitations by utilizing PointNet encoder (Qi et al., 2017) to feature for each point, introducing U-Net-like (Ronneberger, et al., 2015) architecture for convolutional operations, and then Occupancy Network (Mescheder et al., 2019) is used to decode the occupancy probability of each point for 3D reconstruction. Stucker et al. (2022) proposed IMPLICITY that can implicitly generate city-scale DSM, basically, it utilizes the CON network architecture to investigate large-scale 3D reconstruction together with satellite imagery. The generated DSM (Digital Surface Model) can preserve visible details from the original high-resolution UAV imagery, however, due to insufficient attention to local feature structures during the reconstruction process, and an excessive focus on buildings while neglecting non-building features, the Digital Surface Model (DSM) exhibits certain local deformations.

The traditional 3D reconstruction method based on feature point matching is slow, memory-intensive, and often results in noisy reconstructions. Moreover, high-frequency details visible in the images are often lost in the reconstruction results, making it challenging to meet the demands for high-precision geographical information extraction. For deep explicit methods, point cloud representation fails to capture topology and generate watertight surfaces, leading to inferior representation of complex models with intricate topology. Deep implicit methods utilize Signed Distance Function or Occupancy Function to describe the geometric relationship between 3D points and the scene or target surface. Both Signed Distance Function and Occupancy Function are infinitely continuous in 3D space, theoretically possessing infinite resolution, thereby enabling a more precise representation of the scene. This paper addresses on generating Fine-Grained DSM using high-resolution UAV imagery based on the Convolutional Occupancy Networks framework (Peng et al., 2020) and PointNet++ encoder (Qi et al., 2017).

3. Methodology

The goal of this work is to utilize high-resolution unmanned aerial vehicle (UAV) imagery to generate fine-grained Digital Surface Models (DSMs) with sparse or noisy point cloud. An overview of our method is presented in Fig. 2. The proposed Deep-FG-DSM takes the point cloud and ortho-images as input and encodes them into a shape embedding and an image embedding. For any point in 3D space, a corresponding point feature and image feature are interpolated from these two embeddings based on its coordinates. Taking these two features together with coordinates, occupancy decoder is trained to predict the corresponding occupancy probability value. Based on the estimated occupancy probability, two classical methods of Multi-resolution Iso-Surface Extraction and the Marching Cubes (Lorensen et al., 1998) is employed to generate DSMs and reconstruct 3D mesh models, respectively.

In the next subsections, we first provide a more detailed introduction to our network architecture (Section 3.1) and explain the training and inference detail in Section 3.2. Then, the methods for generating DSM and 3D mesh models are introduced (Section 3.3). Final, several variants are discussed in section 3.4.

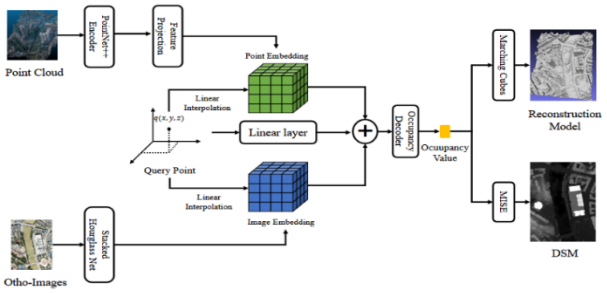


Figure 2. Deep-FG-DSM architecture overview.

3.1 Network Architecture

Inspired by the idea of IMPLICITY (Stucker et al., 2022), as Fig. 2 shows, our Deep-FG-DSM is composed of two feature embedding branches and one occupancy decode head. In particular, PointNet++ proposed by Qi et al. (2017) is applied as encoder to extract feature embedding for point clouds, and we employ the pixel-aligned encoder proposed by Saito et al. (2019) to extract features from RGB ortho-images. Regarding the decoder, we employ a fully connected neural network from the convolutional single-plane decoder which is similar to Peng et al. (2020). Next, more details of each component are introduced as follows:

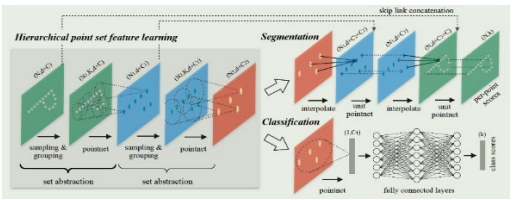


Figure 3. The architecture of the PointNet++ (Qi et al. (2017).

Point Encoder. In order to generate Digital Surface Models (DSMs) that can better capture the local detailed structures of real objects and preserve the fine details of terrain features, we adopt select PointNet++ which is a hierarchical network with density adaptive PointNet layer as point encoder, along with the U-Net encoder (Ronneberger et al., 2015) to extract features from point clouds. PointNet++ recursively applies PointNet to nested partitions of the input point set. The network structure of PointNet++ is illustrated in Figure 3, it performs multi-scale grouping (MSG) on point clouds at different scales based on spatial distance, then extracts features at different scales and group them to form multi-scale features. This enables learning of local features with larger scales. Finally, skip connection is utilized to aggregate feature information across different scales.

According to Peng et al. (2020), we project the d-dimensional features extracted by PointNet++ for each point onto a horizontal plane aligned with the axes of the coordinate frame, discretized at a resolution of H × W pixel cells. This results in planar features with a dimensionality of H × W × d. Subsequently, we employ the U-Net encoder to further process these planar features and yield the point embedding.

Image Encoder. Given the fact that point cloud typically represents limited 3D space which is hard to retain high-frequency information where there is no point, whereas, the high-frequency information is typically clearly presented on high-resolution. We utilize high-resolution orthorectified UAV imagery to extract image features containing high-frequency details, guiding occupancy probability prediction and enhancing the our deep-FG-DSM’s capability to retain original UAV image's high-frequency details. Our method adapts the stacked

hourglass architecture (Newell et al., 2016) employed in PIFu (Saito et al., 2019) as the image encoder for feature extraction. The modified stacked hourglass architecture takes RGB images with rich texture information as input. Furthermore, the features output by this architecture (image embeddings) are dimensionally consistent with the point embeddings and spatially aligned. This alignment enables seamless integration of 3D point cloud and 3D image features corresponding to the input for the decoder, thus predicting the occupancy probability at that point.

Occupancy Decoder. Our method modifies the fully connected network from the convolutional single-plane decoder proposed by Peng et al. (2020) to adapt itself to the dimensions of our input, which is acted as our occupancy decoder. The task of the occupancy decoder is to estimate the occupancy probability at any given position in 3D scene space. For any point $x \in \mathbb{R}^3$, it is projected onto the horizontal (x, y) coordinate plane. Based on its plane coordinates, linear interpolation is utilized to retrieve its corresponding 3D point feature and image feature from the point embedding and image embedding, respectively. These features are then added together and used as the input for the decoder to estimate the occupancy probability at that specific point x.

3.2 Training and Inference

To conduct a feasible training, we combine uniform sampling with shape-adaptive sampling to mitigate overfitting and underfitting. The uniform sampling point resolution is set at 3m, while the surface sampling point resolution is set at 0.4m, and Gaussian noise with σ=0.4m is added to the surface sampling points. These surface sampling points and spatial sampling points constitute query points $\{x_i \in \mathbb{R}^3\}$, with a ratio of approximately 2:1. The training process is supervised by the binary cross-entropy loss $\boldsymbol{L}$ between the predicted occupancy probability $\hat{o}$ and the true occupancy probability o at these query points.

$$\boldsymbol{L}(\boldsymbol{o_i}, \hat{\boldsymbol{o}}_i) = \sum_i (\boldsymbol{o_i} \cdot \mathbf{log}(\hat{\boldsymbol{o}}_i) + (\mathbf{1} - \boldsymbol{o_i}) \cdot \mathbf{log}(\mathbf{1} - \hat{\boldsymbol{o}}_i)) . \quad (1)$$

During model training, the true occupancy probability is derived from the high-precision and reliable Digital Surface Models (DSM) created by a commercial software, Context Capture (CC) within the training area. During inference, for any point $x \in \mathbb{R}^3$ in space, our method predicts the occupancy probability of that point. If the predicted probability is greater than or equal to 0.5, it is considered as an occupied point and assigned an occupancy value of 1; if the probability is less than 0.5, it is considered unoccupied and assigned an occupancy value of 0. Theoretically, our method can generate a DSM with infinite sampling resolution. In practice, when generating the DSM in this work, we first discretize the region of interest into a grid with a horizontal resolution of 0.25m and an initial vertical resolution of 16m. Then, we predict the occupancy value of each grid point. Subsequently, in the vertical direction, we employ the MISE algorithm (Mescheder et al., 2019) in a hierarchical manner to sample and increase the vertical resolution (as detailed in Section 3.3). Ultimately, this process results in a high-precision DSM with a horizontal resolution of 0.25m.

3.3 Generation of DSM and Mesh Model

The primary product of our approach is to generate fine-grained DSMs. Based on the predicted occupancy probabilities, the positions that are supposed to be occupied by 3D points can be derived, we then employ the Multi-resolution Iso-Surface Extraction algorithm (MISE) proposed by Mescheder et al. (2019) to generate the DSMs. As depicted in Fig. 4, the MISE involves conducting four rounds of dense sampling between occupied and unoccupied points in the vertical direction (i.e., along the surface of objects). During each sampling round, two additional points are uniformly inserted between the highest occupied point and its adjacent unoccupied point in the vertical direction, and their occupancy probabilities are computed. This process is iterated four times, continuously interpolating between new occupied and unoccupied points. Theoretically, this approach can enhance the vertical resolution by a factor of 256. Consequently, the generated traditional raster DSM with a horizontal grid spacing of 0.25 meters exhibits higher accuracy.

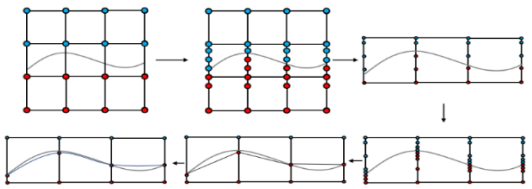


Figure 4. Schematic representation of iso-surface extraction along the vertical direction.

In addition to DSM generation, our method also can facilitate the extraction of 3D mesh models of features directly by utilizing the occupancy probability via the Marching Cubes algorithm proposed by Lorensen et al. (1998). This allows for qualitative analysis of our training results using 3D mesh models.

3.4 Network Variants

Based on the deployed Point Embedding, our network has two variants: the first variant employs the PointNet, while the second variant employs the PointNet++. In the subsequent experimental section, we compare these two embeddings with our method utilizing the PointNet++ encoder and U-Net encoder via using identical datasets and training settings, to investigate the performance of generated fine-grained DSMs.

4. Experiments

Two tests are conducted to evaluate our method. The first one is to investigate the robustness of the proposed Deep-FG-DSM with point cloud of various sparsity and noise. We trained our model using point clouds of different sparsity sampled by various resolutions and different Gaussian noise. The second one is a comparative study. To further demonstrate the capabilities of our model, we compared our method and its variants to PIFU, Convolutional Occupancy Network and IMPLICTY. Our models are implemented by PyTorch (Paszke et al., 2019) on PyCharm. The training process is conducted on a computer equipped with an NVIDIA Quadro P5000 GPU. We employ the Adam optimizer (Kingma et al., 2014) with a decay rate parameter, gamma=0.9999178.

4.1 Dataset and Preprocessing

To the best of our knowledge, only IMPLICTY (Stucker et al., 2022) proposed a training dataset for learning DSM which is from satellite images and not public open, and there is no benchmark of high-resolution UAV images for learning DSM in an implicit manner. To validate the efficacy of our deep-FG-DSM and make community that are interested in dealing with UAV images be able to access relevant benchmark, we generated an aerial image implicit reconstruction dataset [2], named as

[2] More details related to WHU MVS/Stereo dataset will be found at http://gpcv.whu.edu.cn/data/WHU_MVS_Stereo_dataset.html.

And information of our UAVIIR and code can be found at https://github.com/3241674469/Deep-FG-DSM.

UAVIIR (Unmanned Aerial Vehicle Imagery Implicit Reconstruction), it utilizes aerial images from the WHU MVS/Stereo dataset.

Introduction of images and study area. As mentioned, the images of UAVIIR are from the WHU MVS/Stereo dataset, which is originally used for large-scale and high-resolution Multi-View Stereo (MVS) reconstruction. The whole study area within the dataset covers approximately 6.7×2.2 km^2 in Meidan County, Guizhou Province, China. This county comprises densely populated residential and high-rise buildings, sparse factories, forested mountain ranges, as well as some exposed ground and rivers. Images are captured at an altitude of above 550 meters, with a ground sample distance (GSD) of approximately 10 cm. In total, 1,776 images (5376×5376) were acquired across 16 flight strips, providing 90% forward overlap and 80% side overlap. We divided the entire study area into 8 strips, as shown in Fig.5, the size of each strip is approximately 2.2km × 0.8km. Strips 2, 4, and 7 are used as training set (containing varying degrees of land features, including buildings, vegetation, and water bodies which are expected to enhance the model's generalization capability), strip 6 serves as validation set, and strip 5 is designated as the test set. The remaining strips 1, 3, and 8 is also used as test sets to assess the model's generalization capability.

Figure 5. Partition illustration of the WHU MVS/Stereo dataset.

Input Point Cloud and Otho-imagery Generation. The popular commercial software ContextCapture (CC) is used for processing and modelling our aerial images. Initially, high-resolution aerial images were imported, followed by preprocessing steps such as image registration, noise removal, and radiometric correction. Subsequently, aerial triangulation was performed to generate point clouds, ortho-images, precise 3D mesh models and Digital Surface Models (DSM) can be automatically generated. During this procedure, point clouds were obtained, and then simplifying and random Gaussian noise addition were applied to generate sparse point clouds and noisy point clouds, respectively. The UAVIIR dataset includes four types of point clouds with different sampling resolution and four types of point clouds with different levels of Gaussian noise added, as shown in Figure.6.

Query Points Generation. To train our model, it's essential to establish the relationship between any point in space and the target surface. Points inside the target surface are assigned an occupancy value of 1, while points outside the surface are assigned a value of 0. Based on this, we sample query points with known occupancy values in the space of the 3D mesh model generated by CC to supervise our training. However, uniform sampling of points in space may result in most points being far from the target surface, potentially introducing unnecessary external influence on the network. Conversely, sampling only near the iso-surface may lead to overfitting issues. Therefore, we adopt a combined approach of uniform sampling and surface sampling. Specifically, we first randomly sample points on the target surface and add zero-mean Gaussian random noise with σ=0.4m. Then, we perform uniform sampling in space. Specifically, we set the resolution of surface sampling to 0.4m and the resolution of uniform spatial sampling to 3m. The ratio of surface sampling points to uniform spatial sampling points is approximately 2:1. This sampling method effectively preserves information about the target surface while avoiding overfitting issues.

To better evaluate our model's ability to capture fine structural details, we use the detailed structure of buildings in the generated DSM as a criterion. To enhance the generation of fine structural details of buildings, we assign category labels to query points based on the mask images. The mask images are created based on the ortho-image generated by CC, through supervised classification of the ortho-image to distinguish building, forest, and water body categories, followed by creating masks for buildings, forests, and water bodies, respectively. Based on the planar coordinates of the query points, their categories in space are determined. During training, weights are increased for query points belonging to buildings.

$$\boldsymbol{L_t = \lambda L(o_{nb}, \hat{o}_{nb}) + L(o_b, \hat{o}_b)} \quad (2)$$

where, $\boldsymbol{L(o_{nb}, \hat{o}_{nb})}$ and $\boldsymbol{L(o_b, \hat{o}_b)}$ are the corresponding loss for non-building and building query points, respectively, both can be estimated by equation (1). $\boldsymbol{\lambda}$ is a weighting parameter that is set as 0.5 in our work.

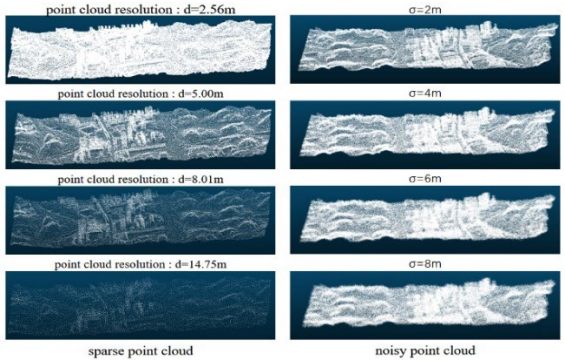


Figure 6. Illustration of the input point cloud.

4.2 Comparison baselines

To demonstrate the efficacy of the proposed deep-FG-DSM, several baseline methods are compared, including various variants discussed in section 3.4 and three relevant learning-based methods whose details are given as follows:

PIFU. The Pixel-aligned Implicit Function (PIFu) method, proposed by Satio et al. (2019), is a widely-used approach for deep implicit surface reconstruction using images. In our experiment, we modified the architecture of the PIFu network via utilizing solely its surface reconstruction component to predict occupancy probability, and the ortho-image that is input by our method is applied.

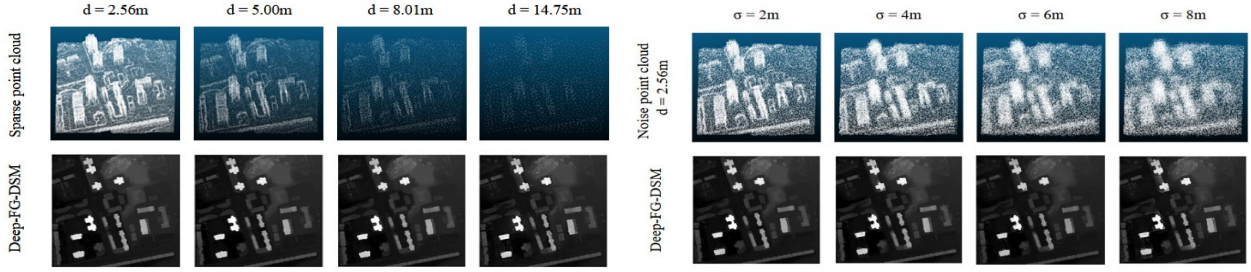
Convolutional Occupancy Network (CON). To compare with CON, in our experiment, we employ Convolutional Single-Plane and Convolutional Volume solutions for occupancy probability prediction.

IMPLICITY. One of the most relevant methods with our deep-FG-DSM is IMPLICITY (Stucker et al., 2022). However, currently, IMPLICITY was only tested on satellite images on which the details of buildings were significantly improved. In our experiment, based on our UAVIIR, we test the IMPLICITY-Mono mode (utilizing only single ortho-image) for occupancy probability prediction.

It is worth noting that for all baselines, based on the decoding occupancy probabilities, Multi-resolution Iso-Surface Extraction (MISE) algorithm is applied for DSM generation.

4.3 Evaluation metrics

In Section 4.1, it mentions that the commercial software CC is applied and can output DSM using high-resolution UAV images with high precision and reliability. For the testing strips, the corresponding DSM is applied as referenced ground truth to evaluate the performance of our Deep-FG-DSM. In evaluation, we align and clip the referenced DSM and the DSM generated by our method pixel-wise.



(a) Different sparse point clouds. (b) Different noisy point clouds.

Figure 7. Qualitative Comparison of DSMs generated from Different simulated point clouds. (a) depicts a 2D schematic of the DSM generated from point clouds by varying simplifying densities. (b) depicts a 2D schematic of the DSM generated from point clouds with sampling resolution of 2.56m, but varying Gaussian noise ($\sigma \in [2m, 4m, 6m, 8m]$). Both of them illustrate a local region of our generated DSM.

Metrics \ Point Cloud	d = 2.56m	d = 5.00m	d = 8.01m	d = 14.75m	σ= 2m	σ= 4m	σ= 6m	σ= 8m
MAE (m)	1.68	2.15	2.35	3.20	2.05	2.14	2.66	3.14
RMSE (m)	3.40	4.41	4.45	5.77	4.08	4.19	5.04	6.01
MedAE (m)	1.09	1.21	1.55	2.04	1.39	1.45	1.79	2.14

Table 1. Quantitative evaluation of our model's performance in generating fine-grained DSMs using different input point clouds.

The discrepancy between the reference height h and the predicted height $\hat{h}$ for each pixel, whereby several metrics of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Median Absolute Error (MedAE) are investigated.

4.4 Ablation study of various point cloud sparsity and noise

In this section, we explored two ablation studies, where the models were trained using the same query points $\{x_i \in R^3\}$. Firstly, we evaluate the robustness of our method with respect to the sparsity of input point cloud. We employ four different sparse point clouds obtained by the Douglas-Peucker algorithm (Douglas et al., 1973), with sampling resolution of 2.56m, 5.00m, 8.01, and 14.75m, for model training and testing. Additionally, to investigate the robustness against noise, based on the point cloud of 2.56m sampling resolution, we simulated four sets of point clouds with Gaussian noise ($\sigma \in [2m, 4m, 6m, 8m]$) for model training and testing.

Sparsity Robustness Study. Fig. 7(a) visually illustrates a local area of DSMs generated by our model using point clouds with different sparsity. This qualitative result reveals that our model exhibits slight performance fluctuations in generating DSMs in response to changes of point cloud sampling density, can effectively capture the edges of objects. Tab. 1 presents the quantitative metrics of DSMs generated by our model, it can be observed that, the accuracy is typically decreased as point cloud sampling resolution grows but with just small magnitude, except for excessively sparse point clouds (e.g., average point cloud resolution d=14.75m). In summary, our model demonstrates robustness against a certain degree of sparse point cloud while still preserves fine details of building boundaries.

Noise Robustness Study. Fig. (b) is a local area of DSMs generated by our model using point clouds with sampling resolution of d=2.56m and varying levels of Gaussian noise. This qualitative analysis shows that our model exhibits minimal performance in generating DSMs when adding various Gaussian noise, and can also effectively preserve the edges of objects. Tab. 1 also provide the accuracy metrics of DSMs generated by our model using point clouds. It can be seen that, our method is able to resist noise to a certain degree.

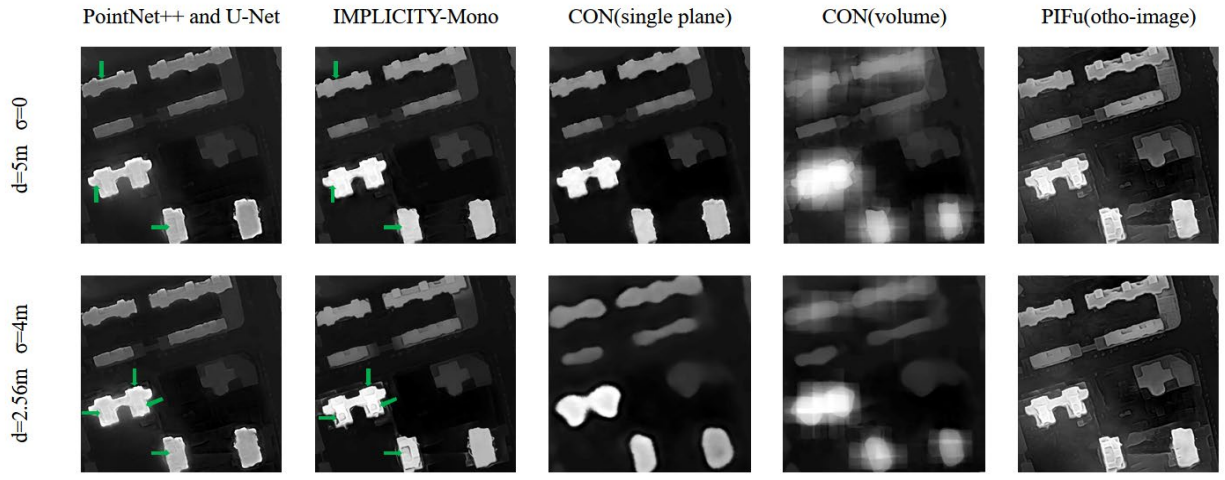


Figure 8. Visual comparison of Deep-FG-DSM with selected baselines.

Metrics \ Point Cloud		PointNet	PointNet++	PointNet++ and U-Net	IMPLICITY -Mono	CON(single plane)	CON(volume)	PIFu(otho-images)
d=5m σ=0	MAE (m)	3.95	4.89	2.15	0.96	1.04	3.22	4.35
	RMSE (m)	6.73	8.14	4.41	3.02	3.48	8.53	7.23
	MedAE (m)	2.30	2.88	1.21	0.35	0.36	0.81	2.61
d=2.56m σ=4m	MAE (m)	4.59	5.37	2.14	1.37	2.64	3.37	4.35
	RMSE (m)	7.56	8.76	4.19	3.66	6.12	7.77	7.23
	MedAE (m)	2.76	3.21	1.45	0.77	1.39	1.31	2.61

Table 2. Quantitative comparison of Deep-FG-DSM and its variants with selected baselines.

4.5 Comparison with other baseline methods.

Several SOTA methods are compared including our own variants. Our method and the relevant variants, along with IMPLICITY-mono, utilize the same point cloud data, query points $\{x_i \in R^3\}$, and ortho-images for both training and testing. The Convolutional Occupancy Networks also employ the same point cloud data and query points $\{x_i \in R^3\}$ for training and testing. Meanwhile, PIFu utilizes the same ortho-images for both training and testing. All methods adhere to the same training strategy and evaluation metrics.

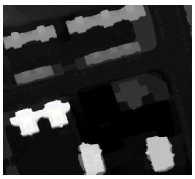


Figure 9. An example region of the DSM generated by CC.

Qualitative Comparison Results. Fig.8 illustrates the visual comparison of 2.5D DSM between our method and other baselines. Variants of Deep-FG-DSM using PointNet and PointNet++ exhibit poor visual effects, their representations are not depicted due to limited space. For the other methods, PIFu, which solely utilizes ortho-image for DSM generation, generate the results that preserve most details and look with sharpest edges, this is mainly due to the fact only ortho-images with rich boundaries are applied, and the relevant accuracy is inferior (see Tab. 2). Our method, Deep-FG-DSM (PointNet++ and U-Net), along with IMPLICITY-Mono, effectively preserves high-frequency details, resulting in visual effects noticeably superior to DSMs generated using Context Capture (CC) (as shown in Fig.9). The DSMs generated by CC are used as reference images for computing evaluation metrics of the generated DSMs. It is clear from the visualization that when using sparse point clouds, our method outperforms IMPLICITY-Mono in terms of DSM visual quality. However, when using noisy dense point clouds, the visual quality of DSMs generated by IMPLICITY surpasses that of Deep-FG-DSM. CON (single-plane) and CON (volume), relying solely on point clouds as input, lack the high-frequency information present on ortho-image, resulting in poor 2.5D visualization effects. PIFU exhibits detailed 2.5D map, but the corresponding mesh model are over-triangulated and contains some mussy meshes, as shown in Fig. 10.

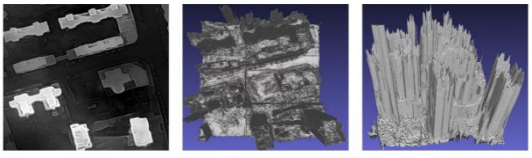


Figure 10. An example region of the 2.5D DSM and the corresponding 3D Mesh model generated by PIFu.

Quantitative Comparison Results. Tab.2 presents a quantitative comparison. While the visual effects of DSMs generated by our method, Deep-FG-DSM, using sparse point clouds, are superior to IMPLICITY-Mono, the accuracy of IMPLICITY-Mono remains higher than our method, this might be explained by the employment of Multi-Scale Grouping (MSG), as depicted in Fig.11, within PointNet++ that is used by Deep-FG-DSM. MSG ensures to learn 3D point feature in the presence of non-uniformly sampled point cloud densities, thereby enhancing the relevant generalization ability regarding various sparsity. However, this also leads to overlapping of partial point cloud information, necessitating pooling operations during feature projection. As a result, the proportionate weight of non-overlapping point cloud information decreases, leading to a decrease in the accuracy of DSM generation by the model. Among all methods, IMPLICITY-Mono achieves the highest accuracy, followed by our method that slightly trails IMPLICITY. Variants of Deep-FG-DSM, due to the absence of the U-Net encoder, exhibit significantly lower accuracy. The Convolutional Occupancy Network (CON) is highly sensitive to noise as it solely relies on point clouds as input, resulting in a severe decrease in DSM accuracy when using large noisy point clouds. PIFu, which only utilizes ortho-image as input, lacks precise spatial information, leading to extremely low accuracy in generated DSMs.

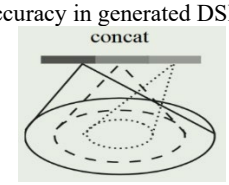


Figure 11. Multi-scale grouping (Qi et al. (2017).

Apart from our method Deep-FG-DSM, the other methods that utilize point clouds exhibit a significant decrease in accuracy when using noisy point clouds, which further demonstrate the fact: the proposed method, Deep-FG-DSM, is capable to resist point cloud noise to a certain degree.

5. Conclusions

In this paper, we propose a deep implicit method, Deep-FG-DSM, for generating fine-grained DSMs with raw noisy and sparse point clouds and ortho-images obtained via photogrammetric processing from high-resolution UAV images. Deep-FG-DSM is capable of handling large-scale scenes and effectively preserving minute details present in real-world environments. The visual quality of DSMs generated by Deep-FG-DSM surpasses those produced by several learning-based methods. Moreover, our Deep-FG-DSM demonstrates better performance when dealing with noisy or sparse point clouds, showing superior robustness to other SOTA deep implicit methods. In the future, we would like to investigate the possibility of using original images instead of ortho-images as more integrated images are very promising to further improve our method's performance. In addition, so far, Deep-FG-DSM has only been trained and tested on our UAVIIR dataset. In the future, we plan to validate the generalization ability of our model on more publicly available high-resolution UAV datasets.

Acknowledgement

This work was jointly supported Natural Science Foundation of Hubei Province, China (2022CFB727) and National Natural Science Foundation of China (42301507).

References

Boykov, Y., & Funka-Lea, G., 2006. Graph cuts and efficient ND image segmentation. *International journal of computer vision, 70*(2), 109-131.

Briechle, K., & Hanebeck, U. D., 2001. Template matching using fast normalized cross correlation. *In Optical Pattern Recognition XII*, Vol. 4387, pp. 95-102. SPIE.

Chen, Z., & Zhang, H., 2019. Learning implicit fields for generative shape modelling. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5939-5948.

Choy, C. B., Xu, D., Gwak, J., Chen, K., & Savarese, S., 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings*, pp. 628-644.

Dabov, K., Foi, A., Katkovnik, V., & Egiazarian, K., 2006. Image denoising with block-matching and 3D filtering. *In Image processing: algorithms and systems, neural networks, and machine learning, Vol. 6064*, *pp. 354-365. SPIE.*

Douglas, D. H., & Peucker, T. K., 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization*, *10*(2), 112-122.

Fan, H., Su, H., & Guibas, L. J., 2017. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 605-613.

Gehrig, S. K., Eberli, F., & Meyer, T., 2009. A real-time low-power stereo vision engine using semi-global matching. In *International Conference on Computer Vision Systems*, pp. 134-143.

Geiger, A., Roser, M., & Urtasun, R., 2010. Efficient large-scale stereo matching. In *Asian conference on computer vision*, pp. 25-38. Berlin, Heidelberg: Springer Berlin Heidelberg.

Gkioxari, G., Malik, J., & Johnson, J., 2019. Mesh r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9785-9795.

Kanade, T., & Okutomi, M., 1994. A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE transactions on pattern analysis and machine intelligence, 16*(9), 920-932.

Kingma, D. P., & Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kanazawa, A., Tulsiani, S., Efros, A. A., & Malik, J., 2018. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision* (ECCV), pp. 371-386.

Lin, C. H., Wang, O., Russell, B. C., Shechtman, E., Kim, V. G., Fisher, M., & Lucey, S., 2019. Photometric mesh optimization for video-aligned 3d object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 969-978.

Lorensen, W. E., & Cline, H. E., 1998. Marching cubes: A high resolution 3D surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pp. 347-353.

Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., & Geiger, A., 2019. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4460-4470.

Michalkiewicz, M., Pontes, J. K., Jack, D., Baktashmotlagh, M., & Eriksson, A., 2019. Implicit surface representations as layers in neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4743-4752.

Newell, A., Yang, K., & Deng, J., 2016. Stacked hourglass networks for human pose estimation. In *Computer Vision – ECCV, Netherlands, Part VIII 14*, pp. 483-499.

Park, J. J., Florence, P., Straub, J., Newcombe, R., & Lovegrove, S., 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165-174.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S., 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., Geiger,A., 2020. Convolutional occupancy networks. *In: Proceedings of the European Conference on Computer Vision.*

Qi, C. R., Su, H., Mo, K., & Guibas, L. J.,2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652-660.

Qi, C. R., Yi, L., Su, H., & Guibas, L. J., 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems, 30.*

Ronneberger, O., Fischer, P., & Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI, Germany, Part III 18*, pp. 234-241.

Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H., 2019. PIFu: Pixel-aligned implicit function for high resolution clothed human digitization. *In: Proceedings of the International Conference on Computer Vision (ICCV)*, 2304-2314.

Scharstein, D., & Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision, 47,* 7-42.

Shen, R., Cheng, I., Li, X., & Basu, A., 2008. Stereo matching using random walks. In *2008 19th International Conference on Pattern Recognition*, pp. 1-4. IEEE.

Stucker, C., Ke, B. X., Yue, Y. W., Huang, S.Y., Armeni, I., Schindler, K., 2022. IMPLICITY: CITY MODELING FROM SATELLITE IMAGES WITH DEEP IMPLICIT OCCUPANCY FIELDS. *ISPRS. Ann. Photogramm. Remote Sens. Spatial Inf. Sci., V-2.*

Wen, C., Zhang, Y., Li, Z., & Fu, Y., 2019. Pixel2mesh++: Multi-view 3d mesh generation via deformation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1042-1051.

Wu, J., Zhang, C., Xue, T., Freeman, B., & Tenenbaum, J., 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modelling. *Advances in neural information processing systems, 29.*

Yang, G., Huang, X., Hao, Z., Liu, M. Y., Belongie, S., & Hariharan, B., 2019. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4541-4550.