DENSE 3D MODEL RECONSTRUCTION FOR DIGITAL CITY USING COMPUTATIONALLY EFFICIENT MULTI-VIEW STEREO NETWORKS

Yuxi Hu$^1$, Zixiao Liu$^1$, Taimeng Fu$^1$ and Man-On Pun$^{1,2,3}$

$^1$School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China, 518172
$^2$Shenzhen Research Institute of Big Data, Shenzhen, China, 518172
$^3$Pengcheng Laboratory, Shenzhen, China, 518055

ABSTRACT

Deep learning has shown promising results on dense three-dimensional (3D) model reconstruction from RGB images in recent years. However, the reconstruction of large-scale 3D models required for digital city remains very challenging even for such deep learning-based methods. In this paper, a convolutional neural network (CNN)-based Multi-View-Stereo (MVS) method is proposed by exploiting a double U-Net approach searching for image features. The proposed network first utilizes a double U-Net to extract the image features of coarser resolution for the sake of reduced memory requirements. After that, the cost volume is built via the differentiable homography warping. The design enables information extraction in a small-scale cost volume before a large-scale cost volume carries out fusion and finer depth map estimation. As a result, the proposed network can efficiently produce highly accurate 3D point clouds using a fraction of the graphics processing unit (GPU) memory and runtime required by conventional methods. Extensive experiments on the DTU benchmarks as well as the Tanks and Temples benchmarks confirm that the proposed network can achieve outstanding reconstruction accuracy and model completeness.

1. INTRODUCTION

Accurate three-dimensional (3D) city models are indispensable for digital city. In recent years, intensive research has been devoted to large-scale 3D model reconstruction in the field of remote sensing [1, 2]. Among the many methods reported in the literature, multi-view stereo (MVS) is one of the most promising approaches. In MVS, a 3D model of a scene is reconstructed from a set of images captured by a camera from different viewpoints with known camera parameters. Indeed, MVS is capable of reconstructing both small objects and large-scale outdoor scenes. However, traditional MVS often fails to obtain accurate reconstruction results in the case of texture starvation, texture repetition, or illumination changes. In addition, MVS is computationally demanding. As a result, it usually takes a long time for MVS to establish the corresponding 3D relationship. This problem is further worsened when reconstructing large-scale scenes. Driven by the rapid development of machine learning technology, the deep learning-based approach has achieved impressive performance on 3D model reconstruction for small objects by mining features and structures of image data. The success in small 3D model reconstruction has hinted that the learning-based MVS can potentially be extended to the 3D reconstruction in large scenes. Some pioneering methods have been proposed by exploiting a feature extraction model to obtain deep features of images before merging these features to build cost volumes for the depth map [3]. In [3], MVSNet was proposed to inherit the theoretical basis of stereo matching while reducing the complexity of the problem by performing depth estimation on only one image at a time. As a result, MVSNet can effectively improve the impact of illumination change problems, which has greatly improved the reconstruction performance in terms of accuracy and generalization. More specifically, images acquired by a sensor with known intrinsic parameters can be translated into cost volumes after differentiable homography warping. After that, 3D convolution operations are performed on the cost volume before the depth map of each and every image is generated by averaging the weighted depth outputs. As a result, 3D point clouds can be obtained by using the photometric and geometric consistencies between multiple images to select the correct depth information for depth prediction. Despite their many advantages, these existing MVS methods are handicapped by their heavy computational requirements and memory consumption. Thus, it is non-trivial to practically deploy these methods to large-scale scenes, particularly when high-resolution image acquisition equipment is employed. Thus, further investigation is required to improve the effectiveness of feature extraction and the regularization of cost volumes in MVS.

Motivated by the aforementioned challenges, this work proposes a multi-scale MVS network by exploiting double U-Net for depth estimation. More specifically, the multi-scale...
feature maps are first generated by the double U-Net feature extraction module. After that, the proposed network performs differentiable homography warping operation on the feature map to warp the feature map of the source image into the reference image before constructing cost volumes. After multiple levels of estimation, the depth map of the original image is obtained before being fused into a point cloud. Extensive experiments on the DTU benchmark are conducted to validate the effectiveness of the proposed network. The main contributions of our approach are summarized as follows: (1) a U-Net architecture is introduced into an end-to-end deep learning based MVS framework for feature extraction enhancement; (2) By casting the model into a coarse-to-fine framework, the hierarchical structure makes the computational complexity sufficiently low for higher resolution images in large-scale scenes.

2. PROPOSED METHOD

Feature Extraction: Feature extraction plays an important role in multi-view stereo matching reconstruction as the reconstruction performance largely depends on the effectiveness of feature extraction. In this paper, an improved double U-Net feature extraction module is proposed as shown in Fig. 1. This module increases the local receptive field for high-frequency regions with rich texture information through two U-Net networks to obtain better feature information. More specifically, given input images of size 512 × 640, Max-pooling and continuous convolution are performed to increase the number of channels from 3 to 8 and then 32, resulting in feature maps of 1, 1/4, and 1/16 of the original image size, respectively. In the upsampling stage, the feature information of different scales is sliced and fused in the channel dimension to form higher-dimensional features. After the concatenations, the convolution and upsampling operations are carried out to obtain a 32-channel feature map of the same resolution as the original image. Next, the output feature map is processed by the U-Net network again to produce three sets of feature maps of different sizes. Such a feature extraction module can retain more detailed features, leading to more accurate depth estimation. It is worth pointing out that the upsampling process is designed to preserve the details and spatial dimensions of the object while side concatenations can better repair the details of the target.

Cost Volume: Most MVS methods use the plane sweeping method reported in [3] to construct the cost volume. Specifically, differentiable homography warping is first performed on the feature map to generate multiple feature volumes. A hierarchical approach to develop the cost volume is more memory efficient as compared to building the cost volume on the entire image. Denote by \( d \) the depth hypothesis, we propose to project all pixels of the source image onto the 3D space using \( d \) followed by reversely warping these projected points by the reference camera. Let \( H_i(d) \) be the homography between the feature maps of the \( i \)-th view and the reference feature map at depth \( d \), we have

\[
H_i(d) = K_i \cdot R_i \cdot \left( I - \frac{(t_i - t_i) \cdot n_1^T}{d} \right) \cdot R_i^T \cdot K_i^{-1},
\]  

(1)

where \( K_i, R_i, t_i \) stand for the camera intrinsics, rotations and translations of the \( i \)-th view, respectively. Furthermore, \( I \) is the identity matrix whereas \( n_1 \) represents the principle axis of the reference camera.

Next, differentiable homography is applied to warp 2D feature maps into hypothesis planes of the reference camera to form feature volumes. To aggregate multiple feature volumes into one cost volume, the variance-based cost metric
is proposed by adopting an arbitrary number of input feature volumes. Mathematically, the cost volume is given by:

\[ C = \frac{1}{N} \sum_{i=1}^{N} (F_i - \bar{F})^2, \]  

(2)

where \( \bar{F} \) refers to the mean of feature volumes while \( N \) is the number of cost volumes.

**Loss function:** The cascaded cost volume with \( N \) stages produces \( N - 1 \) intermediate outputs and a final prediction. The overall weighted loss across all outputs is given by:

\[ \text{Loss}_{\text{total}} = \sum_{k=1}^{N} \lambda^k \cdot L^k, \]  

(3)

where \( L^k \) stands for the smooth L1 loss at the \( k \)-th stage and \( \lambda^k \) refers to its corresponding weighting coefficient.

### 3. EXPERIMENT

**Data description:** The DTU [4] dataset is an indoor dataset specially developed for multi-view stereo tasks. The dataset was generated by an industrial robotic arm equipped with an adjustable brightness light. Images of an object were taken from multiple perspectives. The perspective of each object was strictly controlled while the camera intrinsic and extrinsic parameters were recorded for each view. The dataset consists of 128 different objects or scenes with 79 training scenes, 18 validation scenes and 22 test scenes. Furthermore, each training object or scene was recorded from 49 perspectives each of which has seven brightness levels and 27097 training samples. Each sample is of size 1600 × 1200. For evaluation purposes, the dataset also contains the ground-truth point cloud for each scene. A sample of the dataset is shown in Fig. 2.

The Tanks and Temples dataset contains realistic scenes with small depth ranges. Specifically, its intermediate set is consisted of eight scenes each of which contains hundreds of pictures extracted from high-definition video. In the following section, we will use the DTU training set to train the model while testing the trained model with the DTU evaluation set. To verify the generalization capability of our proposed approach, we will test the trained model on the Tanks and Temples dataset as well as real-life data we measured on the Chinese University of Hong Kong, Shenzhen (CUHKSZ) campus.

**Experimental setup and Results:** Our platform was developed based on PyTorch. Using the DTU training set with all the perspectives and brightness levels, we trained our model with the Adam optimizer for 16 epochs at a learning rate of 0.001. The polynomial decay was adopted during training. Furthermore, the resolution of the DTU input images is 640 × 512 while the number of input images is set to \( N = 5 \). We use a batch size of two to train our model on NVIDIA GeForce RTX 3090. After depth estimation, point clouds were reconstructed as reported in [3, 5]. The model was tested on new perspectives and unknown lighting conditions.

![Fig. 2. Sample experimental data from DTU.](image)

![Fig. 3. Comparison results using the DTU dataset.](image)
real point cloud, which represents the quality of the point cloud reconstructed by MVS. In addition, Comp. calculates the difference between the real point cloud and the reconstructed point cloud, which represents the completeness of point clouds. Overall averages the first two metrics. For all three metrics, a smaller score stands for better performance.

![Fig. 4. Reconstruction point clouds using the DTU dataset.](image)

**Table 1.** Quantitative results on DTU dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc.</th>
<th>Comp.</th>
<th>Overall</th>
<th>GPU(MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gipuma</td>
<td>0.283</td>
<td>0.873</td>
<td>0.578</td>
<td>-</td>
</tr>
<tr>
<td>SurfaceNet</td>
<td>0.450</td>
<td>1.040</td>
<td>0.745</td>
<td>-</td>
</tr>
<tr>
<td>MVSNet</td>
<td>0.456</td>
<td>0.646</td>
<td>0.551</td>
<td>10873</td>
</tr>
<tr>
<td>CasMVSNet</td>
<td>0.365</td>
<td>0.385</td>
<td>0.375</td>
<td>6725</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.350</td>
<td>0.374</td>
<td>0.362</td>
<td>5284</td>
</tr>
</tbody>
</table>

Table 1 provides the quantitative results derived from each method. Inspection of Table 1 suggests that the proposed method outperformed the MVSNet and CasMVSNet in all indices, which implies that the proposed method can reconstruct more accurate points without reconstructing non-existence false points. Note that Gipuma achieved higher accuracy at the cost of higher computational complexity. Thus, Gipuma is impractical for reconstructing large-scale scenes or with higher resolution images.

![Fig. 5. Reconstructed point clouds on Tanks and Temples](image)

Finally, we validate the generalization performance of the proposed method. First, we used the model trained on DTU without any fine-tuning on intermediate dataset of Tanks and Temples. The reconstruction results shown in Fig. 5 look satisfactory. Next, large-scale scenes from the CUHKSZ campus were collected and input into the trained model. From Fig. 6, it is interesting to observe that the proposed method could extract features well for low-textured regions (such as glass) and areas with repeating texture.

![Fig. 6. Real-life large-scale reconstruction at CUHKSZ](image)

### 4. CONCLUSION

Conventional deep learning-based multi-view stereo reconstruction methods are hindered by their ineffective feature extraction and demanding computational complexity. To cope with these problems, a computationally efficient multi-view stereo network with improved feature extraction module has been proposed in this work. The proposed network is featured with a double U-Net structure for feature extraction. Furthermore, a multi-scale cost volume module is developed to increase the connection between cost volumes at different scales through information sharing. As a result, information from different layers can be shared among different cost volumes, which helps to improve the estimation quality of depth maps. Extensive experimental results on the DTU dataset have confirmed that the proposed method can achieve impressive reconstruction accuracy at reduced computational complexity as compared to conventional deep learning-based methods.

### 5. REFERENCES


