M3VSNET: UNSUPERVISED MULTI-METRIC MULTI-VIEW STEREO NETWORK

Baichuan Huang† Hongwei Yi‡ Can Huang† Yijia He† Jingbin Liu† Xiao Liu†

* Wuhan University † Peking University ‡ Megvii Technology Limited

ABSTRACT

The present Multi-view stereo (MVS) methods with supervised learning-based networks have an impressive performance comparing with traditional MVS methods. However, the ground-truth depth maps for training are hard to be obtained and are within limited kinds of scenarios. In this paper, we propose a novel unsupervised multi-metric MVS network, named M3VSNet, for dense point cloud reconstruction without any supervision. To improve the robustness and completeness of point cloud reconstruction, we propose a novel multi-metric loss function that combines pixel-wise and feature-wise loss function to learn the inherent constraints from different perspectives of matching correspondences. Besides, we also incorporate the normal-depth consistency in the 3D point cloud format to improve the accuracy and continuity of the estimated depth maps. Experimental results show that M3VSNet establishes the state-of-the-arts unsupervised method and achieves better performance than previous supervised MVSNet on the DTU dataset and demonstrates the powerful generalization ability on the Tanks & Temples benchmark with effective improvement.

Index Terms— Multi-view stereo, unsupervised, multi-metric, depth map

1. INTRODUCTION

Multi-view stereo (MVS) aims to reconstruct the 3D dense point cloud from multi-view images, which has various applications in augmented reality, virtual reality and robotics, etc. [1, 2]. Big progress has been made in the traditional methods through the hand-crafted features (e.g. NCC) to calculate the matching correspondences [3]. Though, the efficient and robust methods of MVS in the large-scale environments are still the challenging tasks [4]. Recently, deep learning is introduced to relieve this limitation. The supervised learning-based MVS methods achieve significant progress especially improving the efficiency and completeness of dense point cloud reconstruction [5]. These learning-based methods learn and infer the information to handle matching ambiguity which is hard to be obtained by stereo correspondences. However, these supervised learning-based methods strongly depend on the training datasets with ground-truth depth maps, which have limited kinds of scenarios and are not easy to be available. Thus it is a big hurdle and may lead to bad generalization ability in different complex scenarios [6]. Furthermore, the robustness and completeness of dense point cloud reconstruction still have a lot of room to be improved. The learning-based methods are mainly based on the pixel-wise level, which will cause incorrect matching correspondences with low robustness [7]. Because for two identical images, the difference could be huge as long as pixel offset from the perspective of pixel level. However, they are almost the same from the perspective of perception such as feature level. Therefore, the paper aims to the data-independence, robustness and completeness of learning-based MVS.

In this paper, we propose a novel unsupervised multi-metric MVS network, named M3VSNet as shown in figure 1, which could infer the depth maps for dense point cloud reconstruction even in non-ideal environments. Most importantly, we propose a novel multi-metric loss function, namely pixel-wise and feature-wise loss function. The key insight is that the human visual system perceives the surrounding world by the object features. In terms of this loss function, both the photometric and geometric matching consistency can be well guaranteed. Specifically, we introduce the multi-scale feature maps from the pre-trained VGG16 network as vital clues in the feature-wise loss. Low-level feature representations learn more texture details while high-level features learn semantic information with a large receptive field. Different level features are the representations of different receptive fields. Besides, to improve the accuracy and continuity of the depth maps, we incorporate the normal-depth consistency in the world coordinate space to constraint the local surface tangent obtained from the estimated depth maps to be orthogonal to the calculated normal. Therefore, the network can well improve the robustness and accuracy of matching correspondences in some challenging scenarios such as textureless, mirror effect or reflection and texture repeat areas.

2. RELATED WORK

Many traditional methods have been proposed in this field such as voxel-based method [8], feature points diffusion [3] and the fusion of estimated depth maps [9]. The fusion of estimated depth maps can decouple the reconstruction into depth estimation and fusion. Depth estimation with monoc-
Yao Yao proposed MVSNet in 2018 [13], many supervised networks [14, 15, 16, 17] based on MVSNet have been proposed. More importantly, the ground-truth depth maps are not available for all views simultaneously in a symmetric way, which consumes a lot of GPU memory. Additionally, Tejas [18] proposes the simplified network and traditional loss designation but an unsatisfied result. Efforts are worthy to be paid.

3. M³VSNet

3.1. Network Architecture

The basic architecture of M³VSNet consists of three parts, namely pyramid feature aggregation, variance-based cost volume generation and 3D U-Net regularization, as shown in figure 1. The pyramid feature aggregation extracts features from low-level to high-level representations with contextual information. The construction of variance-based cost volume is based on the differentiable homography warping with the number of different depth hypotheses $D$ in MVSNet [13]. At last, the initial depth is derived from the soft argmin operation with the probability volume after the regularization. The advance architecture of M³VSNet consists of normal-depth consistency and multi-metric loss. We incorporate the novel normal-depth consistency to refine depth map in consideration of the orthogonality between normal and local surface tangent. More importantly, we construct multi-metric loss, which consists of pixel-wise loss and feature-wise loss.

3.2. Normal-depth Consistency

The initial depth still contains some incorrect matching correspondences with low quality. Therefore, we incorporate the normal-depth consistency based on the orthogonality between normal and local surface tangent [7]. Due to the orthogonality, the operation of cross-product is used. For each central point $p_i$, one set of the neighbors can be recognized as $p_{ix}$ and $p_{iy}$. If the depth $Z_i$ of $p_i$ and the intrinsics $K$ of camera are known, the normal $\tilde{N}_i$ can be calculated as below:

\[ P_i = K^{-1}Z_ip_i \]  \hspace{1cm} (1)

\[ \tilde{N}_i = P_iP_{ix} \times P_iP_{iy} \]  \hspace{1cm} (2)

The final normal estimation $N_i$ is:

\[ N_i = \frac{1}{8} \sum_{i=1}^{8} (\tilde{N}_i) \]  \hspace{1cm} (3)

In figure 2, for each pixel $p_i(x_i,y_i)$, the depth of the neighbor $p_{\text{neighbor}}$ should be refined. Their corresponding 3D points are $P_i$ and $P_{\text{neighbor}}$. The normal of $P_i$ is $\tilde{N}_i(n_x, n_y, n_z)$. The depth of $P_i$ is $Z_i$ and the depth of $P_{\text{neighbor}}$ is $Z_{\text{neighbor}}$. We can get the equation $\tilde{N}_i \perp P_iP_{\text{neighbor}}$. The relationship is apparently reasonable due to the orthogonality and surface consistency in the local surface.

\[ (K^{-1}Z_iP_i - K^{-1}Z_{\text{neighbor}}P_{\text{neighbor}})(n_x, n_y, n_z) = 0 \]  \hspace{1cm} (4)

Considering the discontinuity of normal in some edge or irregular surface, the weight $w_i$ for the reference image $I_i$ is introduced. The weight is defined as below:

\[ w_i = e^{-\alpha_1|\nabla I_i|} \]  \hspace{1cm} (5)

The weight $w_i$ depends on the gradient between $p_i$ and $p_{\text{neighbor}}$. The final refined depth $\tilde{Z}_{\text{neighbor}}$ is a combination of the weighted sum of different directions.

\[ \tilde{Z}_{\text{neighbor}} = \sum_{i=1}^{8} w'_i Z_{\text{neighbor}} \]  \hspace{1cm} (6)

\[ w'_i = \frac{w_i}{\sum_{i=1}^{8} w_i} \]  \hspace{1cm} (7)

3.3. Multi-metric Loss

We propose a novel multi-metric loss function by considering different perspectives of matching in feature correspondence beyond pixel. The key idea embodied in multi-metric
loss function is the photometric consistency crossing multi-views [9]. Given the reference image \( I_{ref} \) and source image \( I_{src} \), the corresponding intrinsic parameters are represented as \( K_{ref} \) and \( K_{src} \). Besides, the extrinsic from \( I_{ref} \) to \( I_{src} \) is represented as \( T \). For the pixel \( p_i(x_i, y_i) \) in \( I_{ref} \), the corresponding pixel \( p'_i(x'_i, y'_i) \) in \( I_{src} \) can be listed as:

\[
p'_i = K_{src}T(K_{ref}^{-1}Z_i)p_i
\]  

(8)

The overlapping area, named \( I'_{src} \), from \( I_{ref} \) to \( I_{src} \) can be sampled using the bilinear interpolation.

\[
I'_{src} = I_{src}(p'_i)
\]  

(9)

For the occlusion area, the value of the mask \( M \) in \( I'_{src} \) is set to zero. Based on the prior constraint, the multi-metric loss function \( L \) is formulated as the sum of pixel-wise loss \( L_{pixel} \) and feature-wise loss \( L_{feature} \).

\[
L = \sum (\gamma_1 L_{pixel} + \gamma_2 L_{feature})
\]  

(10)

3.3.1. Pixel-wise Loss

For the pixel-wise loss, we only consider the photometric consistency between the reference image \( I_{ref} \) and other source images. There are mainly three parts of this loss function.

Firstly, the photometric loss is:

\[
L_{photo} = \frac{1}{m} \sum\limits_{i=1}^{m} ((I_{ref} - I'_{src}) + (\nabla I_{ref} - \nabla I'_{src})) \cdot M
\]  

(11)

Where \( m \) is the sum number of valid points in the mask \( M \).

Secondly, the loss of structure similarity (SSIM) \( L_{SSIM} \) is:

\[
L_{SSIM} = \frac{1}{m} \sum\limits_{i=1}^{m} 1 - S(I_{ref}, I'_{src})
\]  

(12)

Thirdly, the smooth of final depth map can be operated on the first-order domain and the second-order domain.

\[
L_{smooth} = \frac{1}{n} \sum\limits_{i=1}^{n} (e^{-\alpha_2 |\nabla I'_{src}|} |\nabla Z_i|^2 + e^{-\alpha_3 |\nabla^2 I'_{src}|} |\nabla^2 Z_i|^2)
\]  

(13)

Where \( n \) is the sum number of points in reference image \( I_{ref} \).

Finally, the total pixel-wise loss \( L_{pixel} \) is listed as:

\[
L_{pixel} = \lambda_1 L_{photo} + \lambda_2 L_{SSIM} + \lambda_3 L_{smooth}
\]  

(14)

3.3.2. Feature-wise Loss

The pixel-wise loss performs mismatch errors in some challenging scenarios. Therefore, one of the main improvements of \( M^3VSNet \) is the use of feature-wise loss, which will utilize more semantic information for matching correspondences.

Through the pre-trained VGG16 network, shown in figure 3, the reference image \( I_{ref} \) can extract more semantic high-level information to construct the feature-wise loss function. Here, we extract the layer 8, 15 and 22, which are one half, a quarter and one-eighth the size of the original input images. For every feature from the VGG16, we construct the loss based on the concept of crossing multi-views. Like section 3.3.1, the corresponding pixel \( p'_i \) in \( I_{src} \) can be available. The matching features from \( F_{ref} \) to \( F_{src} \) can be presented as below:

\[
F'_{src} = F_{src}(p'_i)
\]  

(15)

The loss \( L_F \) is represented as below:

\[
L_F = \frac{1}{m} \sum (F_{ref} - F'_{src}) \cdot M
\]  

(16)

The final feature-wise loss function is a weighted sum of different scale of features, which raises the robustness and completeness of point cloud reconstruction. \( L_{F_8} \) represents the feature of layer 8 from pre-trained VGG16.

\[
L_{feature} = \beta_1 L_{F_8} + \beta_2 L_{F_{15}} + \beta_3 L_{F_{22}}
\]  

(17)

4. EXPERIMENTS

4.1. Performance on DTU

The DTU dataset is a multi-view stereo dataset that has 124 different scenes with 49 scans for each scene [19]. With the lighting change, each scan has seven conditions with the known pose. \( M^3VSNet \) is implemented by Pytorch [20]. The resolution of the input image is 640 x 512. Due to the pyramid feature aggregation, the resolution of the final depth is 160 x 128. Additionally, the hypothetical range of depth is sampled from 425mm to 935mm and the depth sample number \( D \) is set to 192. The model is trained with the batchsize as 4 in four NVIDIA RTX 2080Ti. By using adam optimizer for 10 epochs, the learning rates are set to 1e-3 for the first epoch and decrease by 0.5 for every two epochs. For the balance of different weights in loss, we set \( \gamma_1 = 1 \), \( \gamma_2 = 1 \), \( \alpha_1 = 0.1 \), \( \alpha_2 = 0.5 \), \( \alpha_3 = 0.5 \), \( \lambda_1 = 0.8 \), \( \lambda_2 = 0.2 \), \( \lambda_3 = 0.067 \). Beyond that, \( \beta_1 = 0.2 \), \( \beta_2 = 0.8 \), \( \beta_3 = 0.4 \). During each iteration, one reference image and two source images are used. During the testing phase, the resolution of input image is 1600 x 1200.

The official metrics [19] are used to evaluate \( M^3VSNet' \) performance on the DTU dataset. There are three metrics
called accuracy, completeness and overall. As shown in the table 1, M³VSNet outperforms the existed two unsupervised learning-based methods, Unsup_MVS and MVS². Moreover, M³VSNet surpasses the supervised learning-based MVSNet in terms of the overall performance. Compared with traditional MVS methods, M³VSNet achieves significant improvement on the completeness and outperforms Furu and Tola on the overall quality except Colmap but with high efficiency. For more detailed information in point cloud reconstruction, figure 4 illustrates the qualitative comparison. The reconstruction by M³VSNet has more complete texture details than that without feature-wise loss. Therefore, M³VSNet establishes the state-of-the-arts unsupervised learning methods for multi-view stereo reconstruction.

4.2. Generalization Ability on Tanks & Temples

To evaluate the generalization ability of M³VSNet, we use the intermediate Tanks and Temples benchmark that has high-resolution images of outdoor large-scale scenes. The model of M³VSNet trained on the DTU dataset is transferred to the Tanks and Temples benchmark without any finetuning. Another core hyperparameter is the photometric threshold in the process of depth fusion. For the same depth maps, the different photometric thresholds will lead to different performances. Higher photometric threshold will cause better accuracy but worse completeness. In turn, lower photometric threshold will introduce better completeness but worse accuracy. For our proposed M³VSNet, the photometric threshold is set to 0.6 and we get the following results.

The ranking in the Leaderboard of the intermediate Tanks and Temples benchmark shows that M³VSNet is the best unsupervised MVS network until August 30, 2020. What’s more, the DTU dataset is divided into a train-validation-test split. The train-validation-test split has totally different scenes. We train our model in the train split and evaluate the generalization ability by the score in the test split. In view of the above, the performance in figure 5 demonstrates the powerful generalization ability of our proposed M³VSNet.

5. REFERENCES


