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

2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506469

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 M^{3}VSNet, 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 M^{3}VSNet 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 M^{3}VSNet 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-

The code is available at https://github.com/whubaichuan/M3VSNet

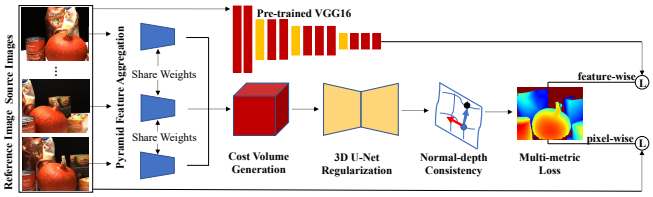

Fig. 1. The architecture of our proposed M^{3}VSNet.

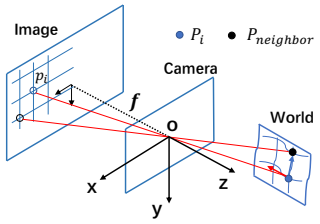

Fig. 2. The illustration of normal-depth consistency

ular video and binocular image pairs has many similarities with the multi-view stereo here [10]. Monocular video [11] lacks the real scale of the depth actually and binocular image pairs always need to rectify the parallel two images [12]. Obstacles such as multi-view occlusion and consistency [6] raise the bar for depth estimation of multi-view stereo than that of monocular video and binocular image pairs. Since 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 derived from heavy labor. Dai [6] predicts the depth maps 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. M^{3}VSNET

3.1. Network Architecture

The basic architecture of M^{3}VSNet 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 M^{3}VSNet 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 $\widetilde{N_i}$ can be calculated as below:

$$P_i = K^{-1} Z_i p_i \tag{1}$$

$$\widetilde{N_i} = \overrightarrow{P_i P_{ix}} \times \overrightarrow{P_i P_{iy}} \tag{2}$$

The final normal estimation N_i is:

$$N_i = \frac{1}{8} \sum_{i=1}^{8} (\widetilde{N_i}) \tag{3}$$

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

$$(K^{-1} Z_i p_i - K^{-1} Z_{neighbor} p_{neighbor})(n_x, n_y, n_z) = 0 \tag{4}$$

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:

$$w_i = e^{-\alpha_1 |\bigtriangledown I_i|} \tag{5}$$

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

$$\widetilde{Z}_{neighbor} = \sum_{i=1}^{8} w_i' Z_{neighbor}^i \tag{6}$$

$$w_i' == \frac{w_i}{\sum_{i=1}^{8} w_i} \tag{7}$$

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:

$$p_i' = K_{src}T(K_{ref}^{-1}\widetilde{Z}_i p_i) \tag{8}$$

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

$$I_{src}' = I_{src}(p_i') \tag{9}$$

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 $L_{feature}$.

$$L = \sum(\gamma_1 L_{pixel} + \gamma_2 L_{feature}) \tag{10}$$

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:

$$L_{photo} = \frac{1}{m}\sum_{i=1}^{m}((I_{ref}^i - I_{src}^{'i}) + (\bigtriangledown I_{ref}^i - \bigtriangledown I_{src}^{'i}))\cdot M \tag{11}$$

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

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

$$L_{SSIM} = \frac{1}{m}\sum_{i=1}^{m}\frac{1 - S(I_{ref}^i, I_{src}^{'i})}{2}\cdot M \tag{12}$$

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

$$L_{smooth} = \frac{1}{n}\sum_{i=1}^{n}(e^{-\alpha_2|\bigtriangledown I_{ref}^i|}\left|\bigtriangledown\widetilde{Z}_i\right| + e^{-\alpha_3|\bigtriangledown^2 I_{ref}^i|}\left|\bigtriangledown^2\widetilde{Z}_i\right|) \tag{13}$$

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

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

$$L_{pixel} = \lambda_1 L_{photo} + \lambda_2 L_{SSIM} + \lambda_3 L_{smooth} \tag{14}$$

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^{3}VSNet 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 F_{src} can be available. The matching features from F_{ref} to F_{src} can be presented as below:

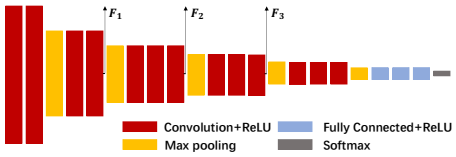

Fig. 3. Feature-wise extraction from pre-trained VGG16

$$F_{src}' = F_{src}(p_i') \tag{15}$$

The loss L_F is represented as below:

$$L_F = \frac{1}{m}\sum(F_{ref} - F_{src}')\cdot M \tag{16}$$

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.

$$L_{feature} = \beta_1 L_{F_8} + \beta_2 L_{F_{15}} + \beta_3 L_{F_{22}} \tag{17}$$

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^{3}VSNet is implemented by Pytorch [20]. The resolution of the input image is 640 $\times$ 512. Due to the pyramid feature aggregation, the resolution of the final depth is 160 $\times$ 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 $\times$ 1200.

The official metrics [19] are used to evaluate M^{3}VSNet' performance on the *DTU* dataset. There are three metrics

Fig. 4. Qualitative comparison on the *DTU* dataset. From left to right: ground truth, MVSNet, M^{3}VSNet without feature-wise loss and M^{3}VSNet.

Method	Mean Distance (mm)		
	Acc.	Comp.	overall.
Furu [3]	0.612	0.939	0.775
Tola [21]	**0.343**	1.190	0.766
Colmap [22]	0.400	**0.664**	**0.532**
SurfaceNet [5]	0.450	1.043	0.746
MVSNet(D=192)	**0.444**	**0.741**	**0.592**
Unsup_MVS [18]	0.881	1.073	0.977
MVS^2 [6]	0.760	**0.515**	0.637
M^{3}VSNet(D=192)	**0.636**	0.531	**0.583**

Table 1. Quantitative results on the *DTU*'s evaluation set. Three classical MVS methods, two supervised learning-based MVS methods and three unsupervised methods using the distance metric (lower is better) are listed.

called accuracy, completeness and overall. As shown in the table 1, M^{3}VSNet outperforms the existed two unsupervised learning-based methods, Unsup_MVS and MVS^2. Moreover, M^{3}VSNet surpasses the supervised learning-based MVSNet in terms of the overall performance. Compared with traditional MVS methods, M^{3}VSNet 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 M^{3}VSNet has more complete texture details than that without feature-wise loss. Therefore, M^{3}VSNet 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 M^{3}VSNet, we use the intermediate *Tanks and Temples* benchmark that has high-resolution images of outdoor large-scale scenes. The model of M^{3}VSNet trained on the *DTU* dataset is transferred to the *Tanks & Temples* benchmark without any finetuning. The intermediate *Tanks and Temples* benchmark contains kinds of

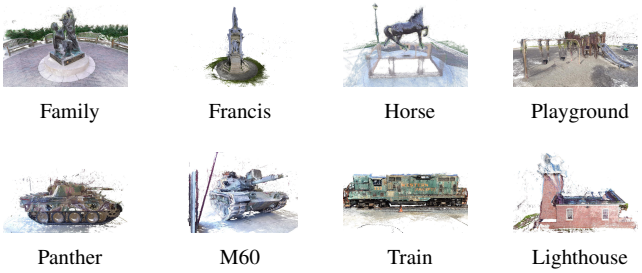

Fig. 5. The performance of M^{3}VSNet on the *Tanks and Temples* benchmark [23] without any finetuning. The quality of dense point cloud reconstruction in large-scale scene shows the powerful generalization ability of M^{3}VSNet.

images with the resolution of 1920×1056 and with the depth hypothesis $D = 160$. 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 M^{3}VSNet, 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 M^{3}VSNet 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 M^{3}VSNet.

5. REFERENCES

[1] Hainan Cui, Xiang Gao, Shuhan Shen, and Zhanyi Hu, "Hsfm: Hybrid structure-from-motion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1212–1221.

[2] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, et al., "Building rome on a cloudless day," in *European Conference on Computer Vision*. Springer, 2010, pp. 368–381.

[3] Yasutaka Furukawa and Jean Ponce, "Accurate, dense, and robust multi-view stereopsis," *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[4] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*. IEEE, 2006, vol. 1, pp. 519–528.

[5] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang, "Surfacenet: An end-to-end 3d neural network for multiview stereopsis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2307–2315.

[6] Yuchao Dai, Zhidong Zhu, Zhibo Rao, and Bo Li, "Mvs2: Deep unsupervised multi-view stereo with multi-view symmetry," in *2019 International Conference on 3D Vision (3DV)*. IEEE, 2019, pp. 1–8.

[7] Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia, "Unsupervised learning of geometry from videos with edge-aware depth-normal consistency," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[8] Sudipta N Sinha, Philippos Mordohai, and Marc Pollefeys, "Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[9] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," in *ACM Transactions on Graphics (ToG)*. ACM, 2009, p. 24.

[10] Hamid Laga, "A survey on deep learning architectures for image-based depth reconstruction," *arXiv preprint arXiv:1906.06113*, 2019.

[11] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1851–1858.

[12] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[13] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 767–783.

[14] Hongwei Yi, Zizhuang Wei, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai, "Pyramid multi-view stereo net with self-adaptive view aggregation," *arXiv preprint arXiv:1912.03001*, 2019.

[15] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo, "P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10452–10461.

[16] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan, "Recurrent mvsnet for high-resolution multi-view stereo depth inference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5525–5534.

[17] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," *arXiv preprint arXiv:1912.06378*, 2019.

[18] Tejas Khot, Shubham Agrawal, Shubham Tulsiani, Christoph Mertz, Simon Lucey, and Martial Hebert, "Learning unsupervised multi-view stereopsis via robust photometric consistency," *arXiv preprint arXiv:1905.02706*, 2019.

[19] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs, "Large scale multi-view stereopsis evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 406–413.

[20] Benoit Steiner, Zachary Devito, Soumith Chintala, Sam Gross, Adam Paszke, Francisco Massa, Adam Lerer, Gregory Chanan, Zeming Lin, Edward Yang, et al., "Pytorch: An imperative style, high-performance deep learning library," pp. 8026–8037, 2019.

[21] Engin Tola, Christoph Strecha, and Pascal Fua, "Efficient large-scale multi-view stereo for ultra high-resolution image sets," *Machine Vision and Applications*, vol. 23, pp. 903–920, 2011.

[22] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision*. Springer, 2016, pp. 501–518.

[23] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–13, 2017.