M^{3}VSNet: Unsupervised Multi-metric Multi-view Stereo Network

Baichuan Huang, Jingbin Liu
Wuhan University, Wuhan, China
huangbaichuan@whu.edu.cn

Can Huang, Yijia He, Xiao Liu
MEGVII, Beijing, China
huangcan@megvii.com

***Abstract*—The present MVS methods with deep learning have an impressive performance than traditional MVS methods. However, the learning-based networks need lots of ground-truth 3D training data, which is not always easy to be available. To relieve the expensive costs, we propose an unsupervised normal-aided multi-metric network, named M^{3}VSNet, for multi-view stereo reconstruction without ground-truth 3D training data. Our network puts forward: (a) Pyramid feature aggregation to extract more contextual information; (b) Normal-depth consistency to make estimated depth maps more reasonable and precise in the real 3D world; (c) The multi-metric combination of pixel-wise and feature-wise loss function to learn the inherent constraint from the perspective of perception beyond the pixel value. The abundant experiments prove our M^{3}VSNet state of the arts in the *DTU* dataset with effective improvement. Without any finetuning, M^{3}VSNet ranks 1_{st} among all unsupervised MVS network on the leaderboard of Tanks & Temples datasets until April 17, 2020. Our codebase is available at https://github.com/whubaichuan/M3VSNet**

I. Introduction

Multi-view stereo (MVS) reconstruction is still a hot topic over the past decade. MVS can be regarded as an extensive process on the basis of structure from motion (SFM) [1] [2]. SFM extracts and matches the feature points from multi-view photos or continuous videos and then reconstructs the sparse point clouds [3] [4]. What's the difference is that MVS aims to reconstruct the dense point clouds [5]. Big progress has been made in the dense construction with traditional methods through the handcrafted algorithm of similarity ((e.g. NCC) and regularization [6] [7] [8] [9] [10]. Though, traditional methods may not work in some scenarios such as textureless, mirror effect, or reflection [5].

To relieve this limitation, deep learning is introduced into MVS [11] [12]. Some outstanding networks, based on convolutional neural network (CNN) and recurrent neural network (RNN), are constructed to infer the information by multi-view stereo correspondences such as MVSNet [13] and R-MVSNet [14]. The features can be learned by the network instead of artificial selection and the inline correspondences can be considered in the forward and backward process, which is proved valid with the constraint of geometric and photometric consistency [15].

The present learning-based MVS methods are very dependent on the ground-truth 3D training data, which is a big hurdle due to the expensive cost for acquiring the training data [16] [13] [17]. One effective solution to that is to construct the unsupervised network without the need for the ground-truth 3D training data [18] [19]. At the same time, deployment and transfer can be easily carried out [20].

The paper introduces a novel method, named M^{3}VSNet, which could infer the depth maps for multi-view stereo without the ground-truth 3D training data. The key insight is derived from that the traditional photometric consistency could be guaranteed based on the correct geometric information [21] [22]. Further, multi-scale information plays a vital role in similarity measurement for textureless regions or no-Lambert surfaces. Previous works such as MVSNet [13] and R-MVSnet [14] extract the feature of only a single layer to construct 3D cost volume, which has been proved to be the important representative for estimated depth [23]. Here, we aggregate multi-scale pyramid features to construct the 3D cost volume with more contextual information. For the loss function of unsupervised methods, previous works [20] [19] [24] pay more attention to the pixel rather than multi-scale features. In view of this, multi-scale feature loss is introduced as a significant supplement. Multi-metric loss including feature-wise loss, which derives from the pre-trained VGG16 network, and pixel-wise loss, can guarantee the understanding in perceptual aspects while the pixel-wise loss focuses on the accuracy of the pixel value. To improve the performance further, normal-depth consistency is introduced to regularize the depth maps in the 3D real space. The regularization will increase the accuracy and precision of depth maps in response to the possible deterioration by multi-scale information.

Our main contributions are summarized as below:

- We propose a novel unsupervised network for multi-view stereo without the ground-truth 3D training data.
- The paper puts forward three methods to deal with textureless regions or no-Lambert surfaces. Multi-metric loss including pixel-wise and feature-wise loss guarantees the understanding in perceptual aspects beyond pixel value. Multi-scale features are extracted to get more contextual information. The normal-depth consistency regularizes the depth maps to be more precise and more reasonable.
- M^{3}VSNet achieves SOTA performance and ranks 1_{st} among all the unsupervised MVS networks on the leaderboard of Tanks & Temples datasets until April 17, 2020.

II. Related Work

A. Traditional MVS

Many traditional methods have been proposed in this field such as voxel-based [25], feature points spread [6], and the fusion of estimated depth maps [26]. The method of voxel-based has to consume many computing resources, whose accuracy depends on the resolution of the voxel [11]. The blank area may suffer from the textureless more serious when feature points diffusion is adopted. The most used method is the fusion of inferred depth maps, which get the depth maps, and then all the depth maps are fused together to output the final point clouds [27]. Many improved insights have been proposed. Neill [27] uses a spatial consistency constraint to remove the outliers from the depth maps. Silvano [9] formulates the patch matches in 3D space and the progress can be massively parallelized and delivered. Johannes [8] estimates the depth and normal maps synchronously, using photometric and geometric priors to refine the image-based depth and normal fusion. Though, the accuracy and completeness can be improved when the problem of the textureless or no-Lambert surfaces can be solved perfectly.

B. Depth estimation

The method of depth maps can decouple the reconstruction into depth estimation and depth fusion. Depth estimation with monocular video and binocular image pair has many similarities with the multi-view stereo here [28]. But there are exactly some differences between them. Monocular video [29] lacks the real scale for the depth actually. Binocular image pairs always rectify the parallel two images [30]. In this case, only the disparity needs to be inferred without considering the intrinsic and extrinsic of the camera. As for multi-view stereo, the input is the arbitrary number of pictures. What's more, the transformation among these cameras should be taken into consideration as a whole [13]. Other obstacles such as multi-view occlusion and consistency [20] raise the bar for multi-view stereo depth estimation than that of monocular video and binocular image pair..

C. Supervised Learing MVS

Since Yao proposes MVSNet in 2018 [13], many supervised networks based on MVSNet have been put forward. To reduce GPU memory consumption, Yao continues to introduce R-MVSNet with the help of GRU [14]. Gu uses the concept of the cascade to shrink the cost volume [16]. Yi introduced two new self-adaptive view aggregation with pyramid multi-scale images to enhance the point clouds in textureless regions [31]. Luo utilizes the plane-sweep volumes with isotropic and anisotropic 3D convolutions to get better results [32]. Yu introduces Fast-MVSNet [17], which firstly gets a sparse cost volume, and then a simple but efficient Gauss-Newton layer can refine the depth maps with great progress inefficiently. In this kind of task, cost volume and 3D regularization are memory consuming and the depth of the true value is derived from heavy labor, which is not fetched easily in other scenarios.

D. Unsupervised Learning in MVS

The unsupervised network utilizes the internal and external constraint to leaning the depth by itself, which relief the complicated and fussy artificial markers for ground-truth depth maps. Many works explore unsupervised learning in monocular video and binocular images with photometric and geometric consistency. Reza [24] presents the unsupervised learning method for depth and ego-motion from monocular video. The paper uses the image reconstruction loss, 3D point cloud alignment loss, and additional image-based loss. Similar to unsupervised learning in monocular video and binocular images [33], the losses of MVS are also the photometric and geometric consistency. Dai [20] predicts the depth maps for all views simultaneously in a symmetric way. In the stage, cross-view photometric and geometric consistency can be guaranteed. But this method consumes a lot of GPU memory. Additionally, Tejas [34] proposes an easy network and traditional loss designation but an unsatisfied result. Efforts are worthy to be paid in this direction.

III. $\mathbf{M^3}$VSNet

In this section, $\mathbf{M^3}$VSNet will be presented in detail. As an unsupervised network, $\mathbf{M^3}$VSNet is based on MVSNet [13]. Our proposed network can work in the multi-view stereo reconstruction without the ground-truth 3D training data, which achieves the best performance among all of the unsupervised MVS networks in accuracy and completeness of point clouds. More importantly, the overall performance of $\mathbf{M^3}$VSNet can be the same as supervised MVSNet in the same setting.

A. Network Architecture

$\mathbf{M^3}$VSNet consists of feature extraction, construction of cost volume, 3D U-Net regularization, normal-depth refine and multi-metric loss. As figure 1 shows, the pyramid feature aggregation with only the finest level is adopted to extract features of the arbitrary number of images. The processes of cost volume, 3D U-Net regularization and initial depth estimation are based on MVSNet, which has been proved effective. Then the initial depth is transferred to the normal domain. In turn, the final depth can be refined with 3D geometric constraint from normal domain to depth domain. Besides, to construct multi-metric loss, another pre-trained network named VGG16 is used to provide the feature-wise constraint. With the traditional pixel-wise constraint, our $\mathbf{M^3}$VSNet can estimate the depth and fuse all of the depth into the final point clouds with the highest level in an unsupervised way.

1) Pyramid Feature Aggregation: In MVSNet [13], only the 1/4 feature is adopted (1/4 represents a quarter of the size of the original reference images). The only one feature map is short of contextual information. There are many choices presented in Lin's work [35]. Featured image pyramid predicts in every different layer and pyramidal feature hierarchy predicts in every hierarchy feature layer. Besides, the feature pyramid network makes the best of contextual information with multi-scale upsampling to predicts independently but with the cost of

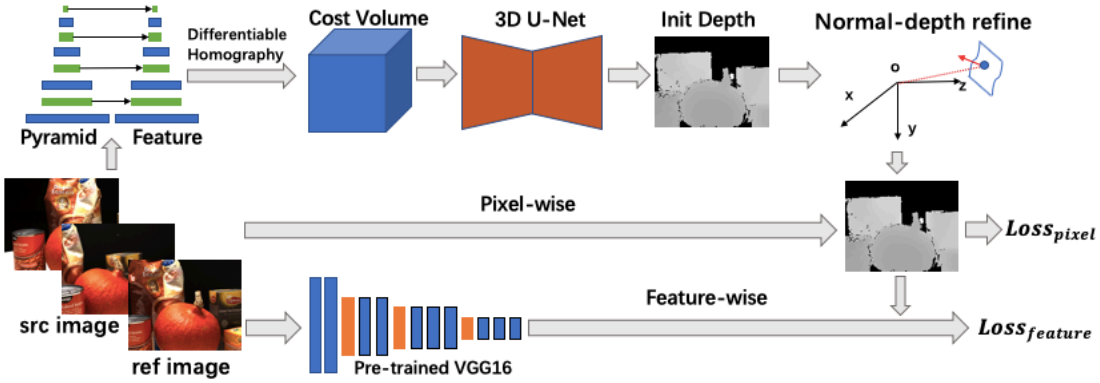


Fig. 1. Our unsupervised network: M^{3}VSNet. It contains four components: pyramid feature aggregation, cost volume and 3D U-Net regularization, normal-depth consistency and multi-metric loss including pixel-wise & feature-wise loss.

more memory consumption. In M^{3}VSNet, the network uses the pyramid feature aggregation with only the finest level, which has been proved helpful than a single feature layer [35]. Next, the aggregation for the pyramid feature will be introduced.

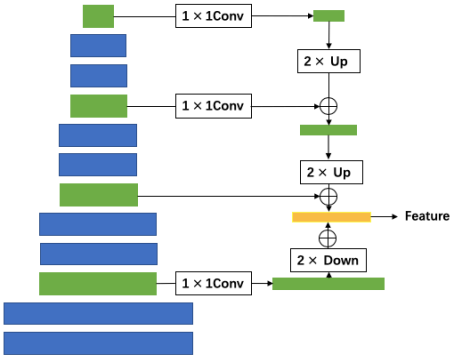


Fig. 2. Pyramid Feature Aggregation

Figure 2 shows the aggregation of pyramid feature. For the input N images, the feature extraction network is constructed to extract the aggregated 1/4 feature. In the process of bottom-up, the stride of 3, 6 and 9 layers are set to 2 to get the four scale features in twelve-layer 2D CNN. Each convolutional layer is followed by the struct of BatchNorm and ReLU. In the process of up-bottom, each level of feature is derived from the concatenate by the upsampling of the higher layer and the feature in the same layer with fewer channels. Especially, the 1/2 feature needs to be downsampling to be aggregated into the final 1/4 feature. To reduce the dimension of the final 1/4 feature, the 1×1 convolution for each concatenation is adopted. At last, we get the final feature with 32 channels, which is an aggregation of contextual information as much as possible.

2) Cost volume and 3D U-Net regularization: The construction of cost volume is based on the homography warping with the different hypotheses of depth [13]. In fact, more depth sampling or fewer depth intervals will lead to better accuracy. Here $D = 192$ is adopted for comparison like the previous two unsupervised methods. Additionally, 3D U-Net regularization can remove the noise by the cost volume, which is the accepted approach for 2D and 3D semantic segmentation. We still use the 3D U-Net in MVSNet, which has simple but effective results. At last, the initial depth is derived from the $soft\ argmin$ operation with the probability volume after regularization. Construction of cost volume and 3D U-Net regularization occupy the most of memory in the whole network. For unsupervised methods, the paper focus on the loss function.

3) Normal-Depth Consistency: The initial depth mainly relies on the probability of feature matches. The textureless and occlusion will lead to the wrong match. How to refine the depth is a key step that can improve the estimated depth. Different from the refine network to the reference image, M^{3}VSNet uses the normal-depth consistency to refine the initial depth in 3D space [19]. The consistency will make the depth more reasonable and accurate. Normal-depth consistency can be divided into two steps. Firstly, the normal should be calculated by the depth with the orthogonality. Then the refined depth can be inferred by the normal and initial depth.

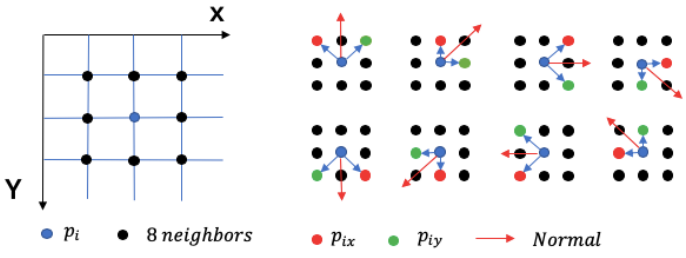


Fig. 3. Normal from the depth

As figure 3 demonstrates, eight neighbors are selected to refer the normal of central point. Due to the orthogonality, the operation of cross-product can be used. For each central point p_i, the match pairs of neighbors can be recognized as p_{ix} and

p_{iy}. If the depth Z_i of p_i and the intrinsics of camera K are known, the normal $\widetilde{N_i}$ can be calculated as below:

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

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

To add the credibility of final normal estimation N_i, mean cross-product for eight neighbors can be presented as below:

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

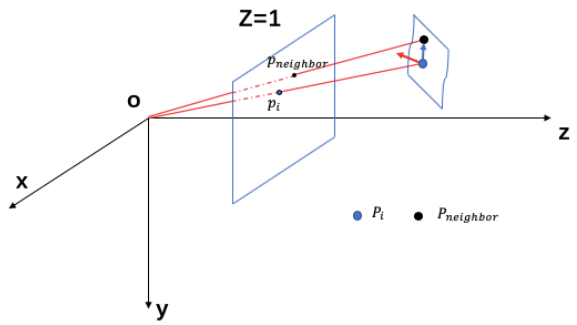


Fig. 4. Depth from the normal

The final refined depth can be available when the normal and initial depth are provided. In figure 4, for each pixel $p_i(x_i, y_i)$, the depth of neighbor $p_{neighbor}$ should be refined. Their corresponding 3D points are P_i and $P_{neighbor}$. Assuming that the normal of P_i is $N_i(n_x, n_y, n_z)$, the depth of P_i is Z_i, the depth of $P_{neighbor}$ is $Z_{neighbor}$, we can get the formula $N \perp P_i \vec{P}_{neighbor}$, which is apparently reasonable due to the orthogonality and surface consistency in the near field. In summary, the depth of neighbors $Z_{neighbor}$ can be inferred by the depth and normal of the central point.

$$(K^{-1} Z_i p_i - K^{-1} Z_{neighbor} p_{neighbor}) \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix} = 0$$

For the refined depth, eight neighbors are also taken into consideration. The neighbors are used to refine the depth of the central point. Considering the discontinuity of normal in some edge or irregular surface of the real object, the weight W_i for the reference image I_i is introduced to make depth more reasonable. The weight is defined as below:

$$w_i = e^{-\alpha_1 |\nabla I_i|}$$

The weight W_i depends on the gradient between p_i and $p_{neighbor}$, which means that the bigger gradient represents the less reliability of the refined depth. In view of the eight neighbors, the final refined depth $\widetilde{Z_i}$ is a combination of weighted sum of eight different directions.

$$\widetilde{Z_i} = \sum_{1}^{8} w_i' Z_{neighbor}$$

$$w_i' == \frac{w_i}{\sum_1^8 w_i}$$

The final refined depth is the results of regularization in 3D space. The 3D geometric constraint makes the depth more accurate and reasonable.

B. Multi-metric Loss

Due to the unsupervised method used here, how to design the loss function is more important. In this paper, the multi-metric loss has played a crucial role. Not only the pixel-wise loss function is introduced, but also the feature-wise loss function is designed to face the disadvantages of textureless and to raise the completeness of point clouds.

The key points embodied in pixel-wise and feature-wise loss function are the photometric consistency crossing multi-views [26]. Given the reference image I_{ref} and source image I_{src}, the corresponding intrinsics K_{ref} and K_{src}, the extrinsic T from I_{ref} to I_{src}. 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 calculated as:

$$p_i' = KT(K^{-1} \widetilde{Z}_i p_i)$$

The overlapping area, named I_{src}', from reference image I_{ref} to source image I_{src} can be sampling using the bilinear method.

$$I_{src}' = I_{src}(p_i')$$

For the occlusion area, the values of pixel in I_{src}' are set to zero. Obviously, the mask M can be obtained when the p_i is projected to the external area of I_{src}. Based on the constraint, the multi-metric loss function L of $\mathbf{M^3}$VSNet can be formulated as the equation.

$$L = \sum (L_{pixel} + L_{feature})$$

1) Pixel-Wise Loss: For the pixel-wise loss, the reference image I_{ref} is used to be the reference to satisfy the consistency crossing multi-views. There are mainly three parts of loss introduced in this section. First, the photometric loss, which compares the difference of pixel value between I_{ref} and I_{src}', is the most used loss. To relieve the influence of lighting changes, the gradient of every pixel is integrated into L_{photo}.

$$L_{photo} = \frac{1}{m} \sum ((I_{ref} - I_{src}') + (\nabla I_{ref} - \nabla I_{src}')) \cdot M$$

Where m is the sum number of valid points according to the mask M

Second, the structure similarity (SSIM) L_{SSIM} is set to measure the similarity of I_{ref} and I_{src}'. The operation S will be 1 when I_{ref} is the same as I_{src}'. The loss function L_{SSIM} aims to make it more similar between I_{ref} and I_{src}'.

$$L_{SSIM} = \frac{1}{m} \sum \frac{1 - S(I_{ref}, I_{src}')}{2} \cdot M$$

Third, the smooth of final refined depth map L_{smooth} can make it less steep in the first-order domain and the second-order domain. Where n is the sum number of points in reference image I_{ref}

$$L_{smooth} = \frac{1}{n}\sum(e^{-\alpha_2|\nabla I_{ref}|}\left|\nabla\widetilde{Z}_i\right| + e^{-\alpha_3|\nabla^2 I_{ref}|}\left|\nabla^2\widetilde{Z}_i\right|)$$

At last, the pixel-wise loss L_{pixel} can be illustrated as below:

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

2) Feature-Wise Loss: Apart from the pixel-wise loss, the main contribution of $\mathbf{M^3}$VSNet is the use of feature-wise loss. For some textureless area, the pixel matching would be wrong, which leads to low precision. But it will be changed with the aid of feature-wise loss. Using more advanced information like high-level semantic information, depth will be well learned even in some textureless regions to some extent.

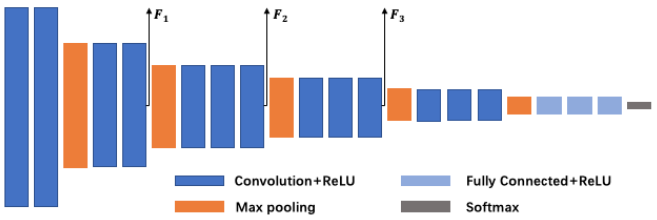


Fig. 5. Feature-wise extraction from pre-trained VGG16

Due to the strong correlation between the estimated depth and pyramid feature network mentioned in section III-A1, the high-level feature is extracted from pre-trained VGG16 instead of the pyramid feature network. By putting the reference image I_{ref} into the pre-trained VGG16 network, showed in figure 5, the feature-wise loss can be constructed. Here, we extract the layer 8, 15 and 22, which are one half, a quarter and one-eighth of the size of the original input images. As a matter of fact, layer 3 output the same size of the original input images, which is actually the reuse of pixel-wise loss.

For every feature from the VGG16, we can construct the loss based on the concept of crossing multi-views. Similar to section III-B1, the corresponding pixel p_i' in F_{src} can be available. The matching feature from F_{ref} to F_{src} can be presented as below:

$$F'_{src} = F_{src}(p'_i)$$

In addition to the pixel value, the feature domain has a bigger receptive field so that the obstacle of textureless regions can be relieved to some extent so that the estimated final depth will detect the similarity of features. The loss L_F is:

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

The final feature-wise loss function is a weighted sum of different scale of features. L_{F_8} corresponds to feature of layer 8 from pre-trained VGG16.

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

IV. Experiments

To prove the effectiveness of our proposed $\mathbf{M^3}$VSNet, this section mainly conducts lots of experiments. First, we explore the performance of $\mathbf{M^3}$VSNet on the DTU dataset including the details of training and testing information. Then the current unsupervised networks in MVS are compared with $\mathbf{M^3}$VSNet. In section IV-C, the ablation studies are carried out to find potential improvement with the proposed contributions. At last, we test $\mathbf{M^3}$VSNet on different datasets such as Tanks and Temples to study the generalization of our model.

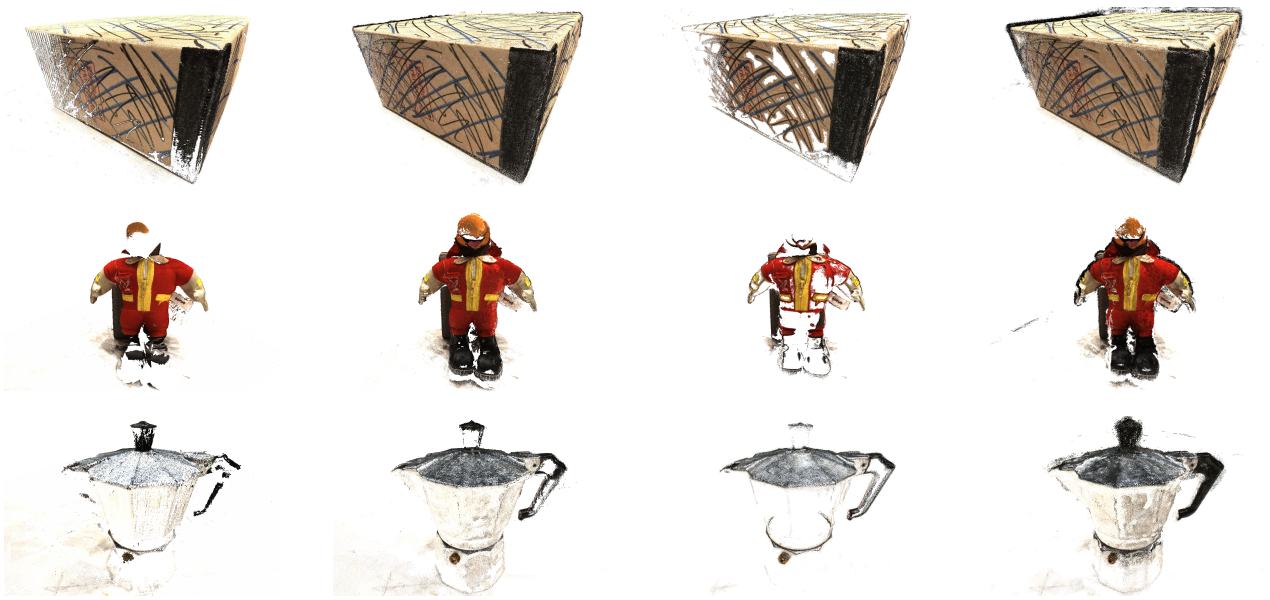
A. Performance on DTU

The DTU dataset is a multi-view stereo set which has 124 different scenes with 49 scans using the robotic arms [36] [37]. By the lighting change, each scan has seven conditions with the pose known. We use the same train-validation-test split as in MVSNet [13] and MVS^2 [20]. Furthermore, the scenes: 1, 4, 9, 10, 11, 12, 13, 15, 23, 24, 29, 32, 33, 34, 48, 49, 62, 75, 77, 110, 114, 118 are selected as the test lists.

1) Implementation detail: $\mathbf{M^3}$VSNet is the unsupervised network based on Pytorch. In the process of training, the DTU's training set without the ground-truth depth maps is used, the resolution of whose is the crop version of the original picture. That is 640 $\times$ 512.Due to the pyramid feature aggregation, the resolution of the final depth is 160 $\times$ 128. The depth ranges are sampled from 425mm to 935mm and the depth sample number is D = 192. The models are trained with the batch of size 4 in four NVidia RTX 2080Ti. By the pattern of data-parallel, each GPU with around 11G available memory could deal with the multi-batch. 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 among loss, we set $\alpha_1 = 0.1$, $\lambda_1 = 0.8$, $\lambda_2 = 0.2$, $\lambda_3 = 0.067$. Additionally, $\beta_1 = 0.2$, $\beta_2 = 0.8$, $\beta_3 = 0.4$. During each iteration, one reference image and two source images are used. In the process of testing, the resolution of input images is 1600 $\times$ 1200, which needs up to 10.612G GPU memory.

2) Results on DTU: The official metrics [36] are used to evaluate $\mathbf{M^3}$VSNet' performance on the DTU dataset. There are three metrics called accuracy, completion and overall. To prove the effectiveness of the model, we compare our proposed $\mathbf{M^3}$VSNet against the three classic traditional methods such as Furu [6], Tola [38] and Colmap [8], and three classic supervised learning methods such as SurfaceNet [11], MVSNet [13] with different depth sample, and the two unsupervised learning methods such as Unsup_MVS [34] and MVS^2 [20].

As table I shows, our proposed $\mathbf{M^3}$VSNet can outperformance the two traditional methods and is so closed to Colmap [8]. As described in MVSNet [13], learning-based methods are



Ground Truth MVSNet Only pixel-wise M^3VSNet

Fig. 6. Qualitative comparison in 3D reconstruction between M^3VSNet and supervised or unsupervised MVS methods on the DTU dataset. From left to right: ground truth, MVSNet(D=256) [13], only the pixel-wise constraint, which is similar to unsup_mvs [34], and our proposed M^3VSNet.

Method	Mean Distance (mm)		
	Acc.	Comp.	overall
Furu [6]	0.612	0.939	0.775
Tola [38]	0.343	1.190	0.766
Colmap [8]	0.400	0.664	0.532
SurfaceNet [11]	0.450	1.043	0.746
MVSNet(D=192)	0.444	0.741	0.592
MVSNet(D=256)	0.396	0.527	0.462
Unsup_MVS [34]	0.881	1.073	0.977
MVS^2 [20]	0.760	0.515	0.637
M^3VSNet(D=192)	**0.636**	**0.531**	**0.583**

TABLE I

QUANTITATIVE RESULTS ON DTUS EVALUATION SET. THREE CLASSICAL MVS METHODS, TWO SUPERVISED LEARNING-BASED MVS METHODS AND TWO UNSUPERVISED METHODS USING THE DISTANCE METRIC (LOWER IS BETTER) ARE LISTED.

ten times more efficient than traditional methods like Colmap. Further, due to the limitation of GPU memory, M^3VSNet selects the sampling value as 192. Obviously, M^3VSNet surpasses the supervised learning method with the same setting. When compared with other unsupervised learning methods, the conclusion can be made that our proposed M^3VSNet is the SOTA network of the unsupervised networks for multi-view stereo reconstruction. For more detailed information in point clouds, figure 6 illustrates the striking contrast. The reconstruction by M^3VSNet has more texture details. With the aid of feature-wise loss and pyramid feature aggregation, M^3VSNet can recover more textureless regions while normal-depth consistency guarantees the accuracy of estimated depth maps in the 3D real space.

B. Comparison with Unsupervised Methods

There are only two unsupervised methods until now. The first one is unsup_mvs [34], which is almost the first try in this direction. But it has poor performance where the overall of mean distance is 0.977. The other method published is MVS^2 [20]. Although MVS^2 can reach to 0.637 in overall of mean distance, it consumes more GPU memory than unsup_mvs due to three cost volumes and regularization needed to be constructed, which is unaffordable for single NVidia RTX 2080Ti used in M^3VSNet. To sum up, our proposed unsupervised method achieves the best performance on the mean distance metric.

C. Ablation Study

The section begins to analyze the effect of different modules proposed in M^3VSNet. There are mainly three contrast experiments carried out. We would explore the role of pyramid feature aggregation, normal-depth consistency and multi-metric loss. All experiments focus on only one variable every time.

Pyramid Feature Aggregation. The module, which can catch more contextual information among different feature layers, is the enhanced version beyond the single feature map. Considering the expensive costs of cost volume construction and 3D U-Net regularization, we use the feature pyramid aggregation with only the 1/4 scale. As table II shows, this module will decrease the value of acc and comp in the mean distance. To summarize, pyramid feature aggregation will improve 2% in overall.

Normal-depth consistency. From an initial depth map to a refined depth map, the module makes the depth map regularized in 3D space, which makes the depth more reasonable.

Method	Mean Distance (mm)		
	Acc.	Comp.	overall
Without Pyramid Feature	0.638	0.554	0.596
With Pyramid Feature	**0.636**	**0.531**	**0.583**

TABLE II
PERFORMANCE COMPARISON WHEN WITH AND WITHOUT THE MODULE OF PYRAMID FEATURE AGGREGATION

Depth error is used to evaluate the quality of estimated depth before the reconstructed point clouds. Here we use the percentage of predicted depths within 2mm, 4mm, and 8mm of ground-truth depth maps (Higher is better). From table III, the performance with the aid of normal-depth consistency surpasses the one without the module in the threshold of 2mm, 4mm and 8mm. Further, in the later step of depth fusion, the contrastive point clouds illustrate the outliers around the object would be removed mostly with the help of normal-depth consistency.

Depth Error (mm)	$\% < 2$	$\% < 4$	$\% < 8$
Without Normal-depth	58.8	74.8	83.8
With Normal-depth	**60.3**	**76.9**	**85.7**

TABLE III
PERFORMANCE COMPARISON WHEN WITH AND WITHOUT THE MODULE OF NORMAL-DEPTH CONSISTENCY

(a) Only pixel-wise (b) No normal-depth (c) Normal-depth

Fig. 7. Qualitative comparison in the reconstruction of 3D point clouds with and without normal-depth consistency

Figure 7 demonstrates the comparison with the case of only pixel-wise loss, with the multi-metric loss but without normal-depth consistency, with both the multi-metric loss and normal-depth consistency. Case (b) and case (c) have better performance than case (a). But apparently, case (b) has more outliers than case (c), which proves that the module of normal-depth consistency can make depth maps more reasonable to some extent but a little deterioration in terms of completeness. The explanation is that the 3D space regularization can guarantee the refined depth maps to follow the rule in the real world. Figure 7 and table III can prove the significant benefits of this module.

Multi-metric loss. The most unsupervised networks about depth estimation, either monocular video or binocular rectified image pairs, focus on the pixel-wise loss construction, which conforms to humans' thoughts. But the constraint pointing to feature-wise is effective in previous related work [39] [40] [18]. We have compared the pixel-wise loss only and the different combinations of feature-wise loss. The multi-metric loss shows a big improvement. What's more, how to select the multi-scale features is also taken into comparison.

As the telling in table IV, the overall of only pixel-wise loss is relatively higher. The different combinations of feature-wise losses make it an impressive improvement. Further, we do some ablation studies on the different scales of features from pre-trained VGG16. The 1/4 feature is matched to the resolution of depth map by the network's output. The results show that the combination of 1/2, 1/4, 1/8 features achieves the best result. By the way, adding the 1/8 feature improves the accuracy but deteriorate the completeness. The cause may be that too advanced semantic information is out of control under the estimated depth.

Method	Mean Distance (mm)		
	Acc.	Comp.	overall
Only pixe-wise	0.832	0.924	0.878
pixel+ 1/4 feature	0.646	0.591	0.618
pixel+ 1/2,1/4,1/8 feature	**0.636**	**0.531**	**0.583**
pixel+ 1/2,1/4,1/8,1/16 feature	0.566	0.653	0.609

TABLE IV
PERFORMANCE COMPARISON OF THE DIFFERENT LOSSES. WHERE THE SCALE OF 1/2 REPRESENTS THAT THE FEATURE (CORRESPONDING TO LAYER 8) EXTRACTED FROM THE PRE-TRAINED VGG16 NETWORKS IS HALF OF THE ORIGINAL REFERENCE IMAGE. THE SCALES OF 1/4, 1/8, 1/16 CORRESPOND TO LAYER 15, 22, 29 IN PRE-TRAINED VGG16.

D. Generalization Ability on Tanks & Temples

To evaluate the generalization ability of our unsupervised network, the intermediate Tanks and Temples dataset, which has high-resolution images of outdoor scenes, is adopted. The models of $\mathbf{M^3}$VSNet trained on the DTU dataset is transferred without any finetuning. We use the intermediate scenes with the resolution of 1920×1056 and 160 depth intervals because 192 depth intervals will out of memory. What's more, another core hyperparameter is the photometric threshold in the process of depth fusion. For the same depth maps of whole datasets, the different photometric thresholds will lead to different performances. In other words, the hyper-parameter will cause the change of accuracy and completeness. For $\mathbf{M^3}$VSNet, the photometric threshold is set to 0.6 and we get the following results. As shown in table V, the ranking is selected from the leaderboard of intermediate T&T. $\mathbf{M^3}$VSNet is better than MVS^2 by the mean score of 8 scenes, which is the best unsupervised MVS network until April 17, 2020. The point clouds are presented in figure 8, which are detailed and reasonable for scenes Family, Francis, Horse, M60, Panther, Playground, Train, Lighthouse. It's worth noting that $\mathbf{M^3}$VSNet can be applied to advanced T&T but the reconstruction is so sparse due to the limitation of GPU memory. It's a balance between GPU memory consumption and the performance of point clouds.

V. CONCLUSION

In this paper, we proposed an unsupervised network for multi-view stereo reconstruction named $\mathbf{M^3}$VSNet, which

Method	Mean	Family	Francis	Horse	Lightouse	M60	Panther	Playground	Train
M^{3}VSNet	**37.67**	**47.74**	**24.38**	**18.74**	**44.42**	**43.45**	**44.95**	**47.39**	**30.31**
MVS^2	37.21	47.74	21.55	19.50	44.54	44.86	46.32	43.48	29.72

TABLE V
QUALITATIVE COMPARISON IN 3D POINT CLOUDS RECONSTRUCTION ON THE TANKS AND TEMPLES DATASET [41] AMONG ALL THE UNSUPERVISED METHODS, WHICH IS FROM THE LEADERBOARD OF INTERMEDIATE T&T.

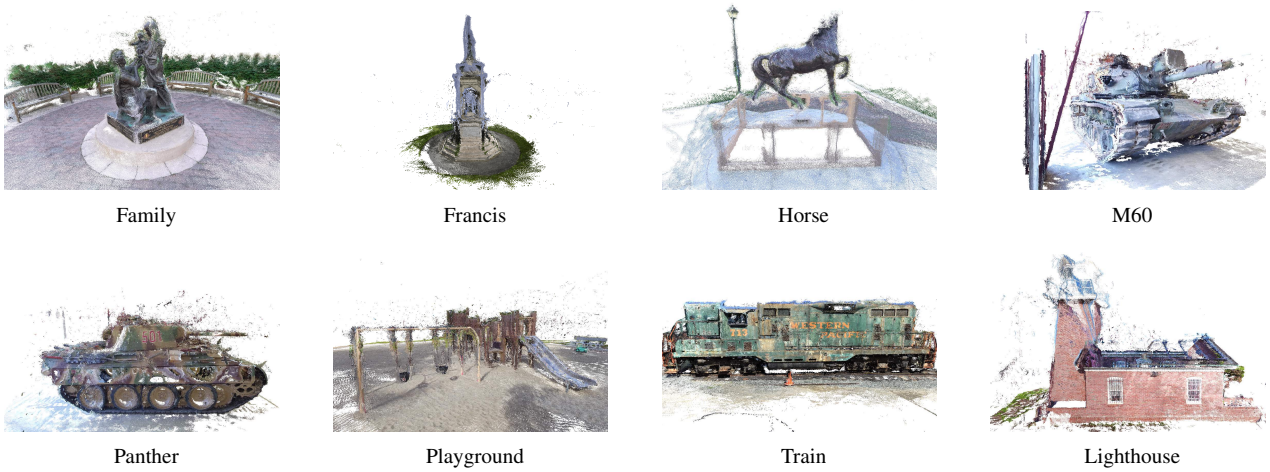


Fig. 8. Our unsupervised network's performance on the Tanks and Temples dataset [41] without any finetuning.

achieves the state of arts in unsupervised MVS networks. With our proposed methods of pyramid feature aggregation, normal-depth consistency and multi-metric loss, M^{3}VSNet can capture contextual and high-level semantic information from the perspective of perception, and make sure the rationality of estimated depth maps in the real 3D world as to make it the best performance on DTU and other MVS datasets among all the unsupervised networks. In the future, more MVS datasets with high precision are desired. Besides, the domain transfer for different datasets can be improved better. Like the prosperity of other works in computer vision with deep learning, multi-task such as semantic, instance segmentation and depth completion can be combined with multi-view stereo reconstruction for the time to come.

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*, pages 1212–1221, 2017.

[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*, pages 368–381. Springer, 2010.

[3] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM Siggraph 2006 Papers*, pages 835–846. 2006.

[4] Roberto Tron, Xiaowei Zhou, and Kostas Daniilidis. A survey on rotation optimization in structure from motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 77–85, 2016.

[5] 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)*, volume 1, pages 519–528. IEEE, 2006.

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

[7] Michael Goesele, Brian Curless, and Steven M Seitz. Multi-view stereo revisited. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2402–2409. IEEE, 2006.

[8] 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*, pages 501–518. Springer, 2016.

[9] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015.

[10] Hoang-Hiep Vu, Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven. High accuracy and visibility-consistent dense multiview stereo. *IEEE transactions on pattern analysis and machine intelligence*, 34(5):889–901, 2011.

[11] 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*, pages 2307–2315, 2017.

[12] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5038–5047, 2017.

[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)*, pages 767–783, 2018.

[14] 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*, pages 5525–5534, 2019.

[15] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-

Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018.

[16] 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.

[17] Zehao Yu and Shenghua Gao. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. *arXiv preprint arXiv:2003.13017*, 2020.

[18] Wang Benzhang, Feng Yiliu, Fu Huini, and Hengzhu Liu. Unsupervised stereo depth estimation refined by perceptual loss. In *2018 Ubiquitous Positioning, Indoor Navigation and Location-Based Services (UPINLBS)*, pages 1–6. IEEE, 2018.

[19] 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.

[20] 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)*, pages 1–8. IEEE, 2019.

[21] Martin Weber, Andrew Blake, and Roberto Cipolla. Towards a complete dense geometric and photometric reconstruction under varying pose and illumination. In *BMVC*, pages 1–10, 2002.

[22] Tianwei Shen, Lei Zhou, Zixin Luo, Yao Yao, Shiwei Li, Jiahui Zhang, Tian Fang, and Long Quan. Self-supervised learning of depth and motion under photometric inconsistency. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[23] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018.

[24] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5667–5675, 2018.

[25] 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*, pages 1–8. IEEE, 2007.

[26] 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)*, page 24. ACM, 2009.

[27] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conference on Computer Vision*, pages 766–779. Springer, 2008.

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

[29] 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*, pages 1851–1858, 2017.

[30] 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*, pages 2758–2766, 2015.

[31] 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.

[32] 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*, pages 10452–10461, 2019.

[33] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941*, 2018.

[34] 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.

[35] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[36] 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*, pages 406–413, 2014.

[37] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120:153–168, 2016.

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

[39] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[40] Anjie Wang, Zhijun Fang, Yongbin Gao, Xiaoyan Jiang, and Siwei Ma. Depth estimation of video sequences with perceptual losses. *IEEE Access*, 6:30536–30546, 2018.

[41] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.