NPF-MVSNet: Normal and Pyramid Feature Aided Unsupervised MVSNet

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—We proposed an unsupervised learning based network, named NPF-MVSNet, for multi-view stereo reconstruction without ground-truth 3D training data. Our network put forward: (a)The pyramid feature aggregation to capture more contextual information for cost volume construction; (b) Normal-depth consistency to make estimated depth maps more reasonable and precise in real 3D world; (c) the combination of pixel-wise and feature-wise loss function to learning the inherent constraint from the perspective of perception beyond the pixel value. The experiments has prove the state of arts of NPF-MVSNet and the each innovation insights contributes to the network with effective improvement. The excellent generalization ability of our network without any finetuning is shows in the leaderboard of Tanks & Temples datasets. NPF-MVSNet is the best unsupervised MVS network with the limited GPU memory consumption until April 17,2020. Our codebase is available at https://github.com/whubaichuan/NPF-MVSNet

I. INTRODUCTION

Mluti-view stereo reconstruction is still a hot topic in recently years. MVS can be regarded as the extend process in the basis of SFM, which extracts and matches the feature points from multi view photos or a continuous videos and then reconstruct the sparse point cloud. What's the different is that MVS aims to reconstruct the dense point cloud. Many traditional method has been proposed in this field such as voxel based, feature points spread and depth estimation. A big progress has been made in the dense construction with traditional methods though the handcrafted algorithm of similarity and regularization have been designed to calculate the dense matches. But the features selected may not work in some scenarios such as textureless, mirror effect or reflection. High accuracy but not high completeness is the common disadvantages which can be improved.

Deep learning has made a great development in image detection and segmentation, language understanding and extend area of CV and NLP. Many works prove that the effective of CNN and RNN. The features can be learned by the network instead of artificial selection and the inline relationship can be considered in the forward and backward process. Reconstructing the dense point cloud using end to end methods has been proved valid with the constraint in geometry and graphics.

Three main ways to multi-view stereo reconstruction are voxel based [9], feature points diffusion and the fusion of depth map [3]. The method of voxel based have to consume many computing resources, whose accuracy depends on the resolution of voxel. Blank area may suffer from the textureless more when feature points diffusion is adopted. Further, textureless is a big problem over the reconstruction for many years. Not only in the traditional computer vison and graphics, but also with the help of neural network the textureless can not be solve perfectly. In industry, human-inhenced features are designed to be projected to the surface of object using the Infrared light spot. Besides this, many improving methods in algorithm have been proposed such as multi scale, semantic information, high dimensional feature and so on. Though, the most used of multi-view stereo reconstruction is the fusion of inferred depth, which get the depth for the selected picture with a pair of images and then all the depth are fused together to output the final points cloud.

Depth estimation with monocular video and binocular image pair has much similarity with the multi-view stereo here [13]. But there are exactly some different between them. Monocular video lack the real scale for the depth actually, which can be estimated by the external help like IMU. Binocular always rectify the parallel two images. In the case, only the disparity need to be inferred without considering the intrinsics and extrinsics of 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. There are many other obstacles such as muti-view occlusion and lightening changes raising the bar for multi-view stereo reconstruction.

The paper introduces a novel method that achieve the state of arts in unsupervised model with the limited GPU memory. There are three main contribution of NPF-MVSNet, which make NPF-MVSNet the best model in unsupervised MVSNet domain until April 17, 2020.

- one is that the feature pyramid with only the finest level is used to extract more contextual information with the consuming memory as little as possible.
- Another promote is that the network utilizes the normal-depth consistency to regularize the estimated depth in 3D space. The normal-depth consistency make the depth more precise and more reasonable.
- The last insight is that the network constructs the feature-wise loss function along with pixel-wise loss function, which derives from the pretrained VGG16 network. The advantages of feature-wise loss can guarantee the understanding in perceptual aspects while the pixel-wise loss

focus on the accuracy of pixel value.

II. RELATED WORK

A. MVS reconstruction

AS talked above, traditional MVS reconstruction can be divide into three main approaches, among which the most used method is the fusion of inferred depth. Neill use a spatial consistency constraint to remove the outliers from the depth maps [3]. Silvano [6] formulate the patchmatch to in 3D space and the progress can be massively parallelized and delivers. Johannes [17] estimate the depth and normal synchronously. Using photometric and geometric priors to refine the image-based depth and normal fusion. The fusion of depth decouple the multi-view stereo reconstruction into two steps. One is depth inferring and the other is fusing all the depth maps, which enable us to pay attention to only one steps by one time. Though, the accuracy and completeness can be improved when the problem of textureless or no-Lambert surface can be solved perfectly.

B. Supervised Learing in MVS

Since Yao proposes MVSNet in 2018 [20], many super-vised networks based on MVSNet have been put forward. To reduced the GPU memory consumption, Yao continues to introduce R-MVSNet with the help of GRU [21]. Gu use the concept of cascade to shrink the cost volumn [7]. Yi introduced two new self-adaptive view aggregation with pyramid multi-scale images to enhance the points cloud in textureless regions [22]. Luo utilizes the plane-sweep volumes with isotropic and anisotropic 3D convolutions to get better results [15]. Yu introduce Fast-MVSNet [23], which first get a sparse cost volume and then a simple but efficient Gauss-Newton layer can refined the depth maps with a great progress in efficient. In such area, cost volume and 3D regularization is memory consuming and the depth of the true value is derived from heavy labor, which is not fetched easily in other scenario.

C. Unsupervised Learning in MVS

Unsupervised network utilizes the internal and external constraint to leaning the depth by itself, which relief the complicate and fussy artificial mark for depth in ground truth. Many works explore the unsupervised learning in monocular video and binocular images with the photometric and geometric consistency. Reza presents the unsupervised learning for depth and ego-motion from monocular video [16]. The paper uses the image reconstruction loss, 3D point cloud alignment loss and additional image-based loss such as photometric and smooth. In the area of unsupervised learning, the design of loss is more important. Similar to unsupervised learning in monocular video and binocular images [1], the composition of loss of MVS is also the photometric and geometric consistency. Until now, there are few unsupervised learning MVS proposed. Dai predicts the depth maps for all views simultaneously in a symmetric way. In the stage, cross-view photometric and geometric consistency can be guaranteed . [4]. But the method consumes so many GPU memory, which is at least three times than MVSNet [20]. Additionally, Tejas proposes a easy network and traditional loss designation [11]. The paper has a unsatisfied results but set a start for unsupervised learning for MVS. Efforts is worthy to be payed in this direction. The lower memory consumption and the more effective loss designation are both emerging topic.

III. NPF-MVSNET

In this section, NPF-MVSNet will be presented in detail. As an unsupervised network, NPF-MVSNet is based on MVSNet [20]. Our proposed network can work in the multi-view stereo reconstruction without the ground truth MVS data and achieves the best performance among all of the unsupervised MVSNet in accuracy and completeness of point clouds with the limited GPU memory. More important, the overall performance of NPF-MVSNet can be the same with supervised MVSNet in the same setting. In our proposed condition, NPF-MVSNet is better than supervised MVSNet.

A. Network Architecture

NPF-MVSNet consists of feature extraction, construction of cost volume, 3D U-Net regularization, normal-depth refine and loss designation. As Figure 1 shows, the feature pyramid with only the finest level is adopted to extract features of inputting arbitrary number of pictures. The processes of cost volume, 3D U-Net regularization and initial depth estimation are much similar as MVSNet, which has been proved as effective. The the initial depth is transferred to the the normal domain. From normal domain to depth domain, the final depth can be refined with 3D geometric constraint. Besides, another pretrained network named VGG16 is used to provide the feature-wise constraint. With the traditional pixel-wise constraint, our NPF-MVSNet can estimate the depth and fuse all of the depth into the final point clouds with the highest level in unsupervised way.

1) Pyramid Feature Aggregation: In MVSNet, only the 1/4 feature is adopted in single feature map. The lower layer of feature has less semantic information and the only one feature map is lack of the contextual information between different layer of feature. There are many choices presented in Lin's work [14]. Featurized image pyramid predicts in every different scale images layer and feature pyramidal feature hierarchy predicts in every hierarchy feature layer. Feature pyramid network make the best of contextual information with upsampling to predicts independently. Although the feature pyramid will get the best results and the idea of multi scale prediction is reasonable and effective, they consumes much memory which is not be allowed in MVSNet. In NPF-MVSNet, the network uses the feature pyramid with only the finest level, which has been proved helpful than single feature map [14]. Next, the aggregation for pyramid feature will be introduced.

Figure 2 shows the aggregation of pyramid feature. For the input N images, the feature pyramid can be constructed to extract the aggregated 1/4 features. In the process of bottom-up, the four scale features map are used in eleven-layer 2D
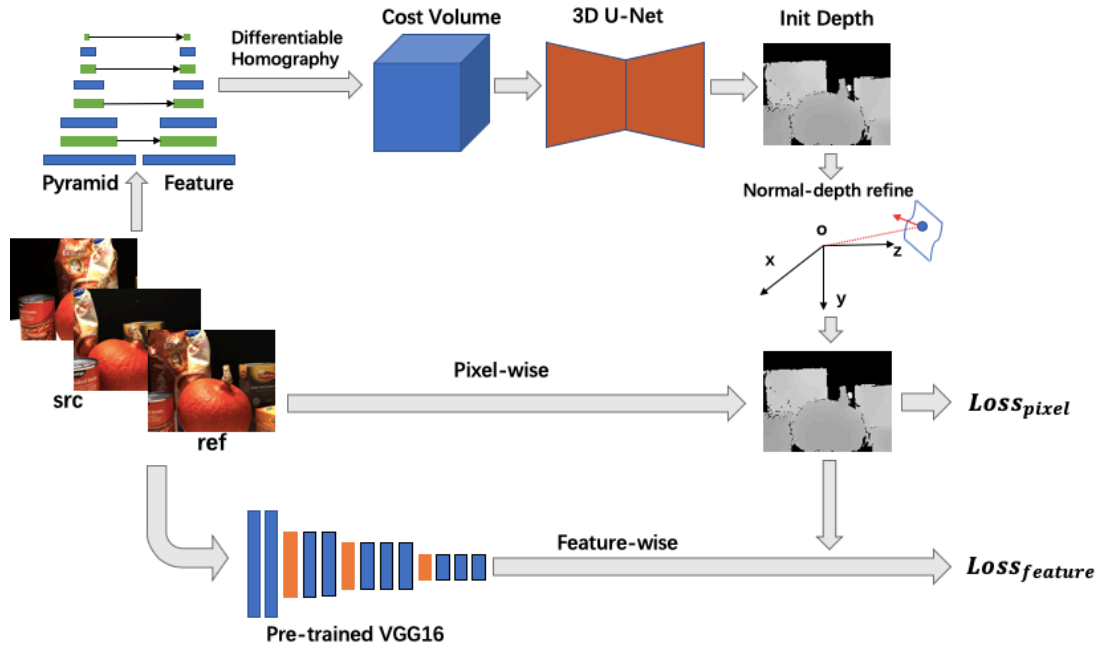
Fig. 1. Our unsupervised network: NPF-MVSNet. It contains four components: pyramid feature aggregation, cost volume and 3D U-Net regularization, normal-depth consistency and pixel-wise & feature-wise loss.
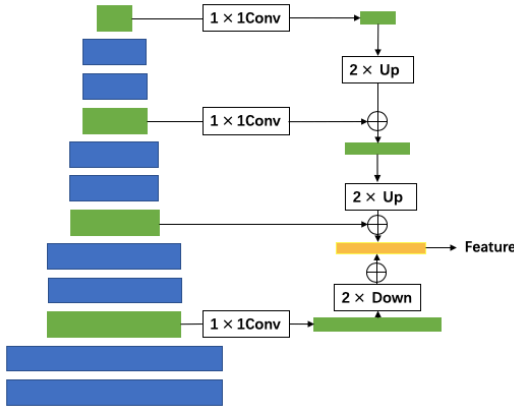

Fig. 2. Pyramid Feature Aggregation

CNN. The stride of 3,6 and 9 are set to 2 to get the four scale features. Each convolutional layer is followed by the struct of BatchNorm and ReLU. In the process of up-bottom, the each level of features are derived from the concatenate by the upsampling of higher layer and the feature in same layer in bottom-up. Specially, the features of 1/2 scale need to be downsampling to be aggregated into final features of 1/4 scale. To reduce the dimension of final feature, for each concatenation, the features in bottom-up are produced by the 1×1 convolution. At last, we get the final features with 32 channels, which is a fusion of different information as a thorough control. The aggregation for pyramid feature can get information in different layer as much as possible with the

limiter GPU memory.

2) Cost volume and 3D U-Net regularization: The process of cost volume is based on the homography warping with the different hypothesis in depth [20]. In fact, the bigger number of depth sampling or the smaller number of depth interval, the more accuracy of the estimated depth. Here $D = 192$ is adopted in MVSNet for comparision like previous two unsupervised method. Based on the setup, the overall is 0.592 for MVSNet [4]. 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 the simple but effective results. Although the structure of 3D U-Net is strengthened in the two unsupervised methods, the improvement in accuracy with the cost of GPU memory. The two process: construction of cost volume and 3D U-Net regularization occupy the most of memory in whole network. What we can conclude is that the only accumulation in depth sampling and the structure of regularization make little sense. For unsupervised methods, the paper focus on the loss function. At last, the depth is derived from the $soft\ argmin$ operation with the probability volume.

3) Normal-Depth Consistency: The initial depth outputted from the regularization mainly relies on the probability of feature matches. The textureless and occlusion will lead to the wrong match so how to refine the depth is a key step which can improve the estimated depth. Different from the refine network with ref images, NPF-MVSNet uses the normal-depth consistency to refine the depth in 3D space. The consistency will make the depth more reasonable and accuracy.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.
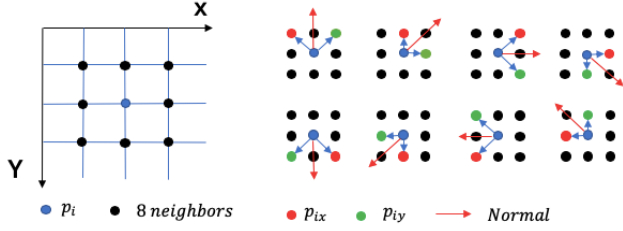

Fig. 3. Normal from the depth

As Figure 3 demonstrated, eight neighbors points are selected to refer the normal of central points. Due to the orthogonality, the operation of cross-product can be used. For each pixel p_i, the match pair 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 can calculated as below:

$$P_i = ZK^{-1}p_i$$

$$N_i = P_i\vec{P}_{ix} \times P_i\vec{P}_{iy}$$

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

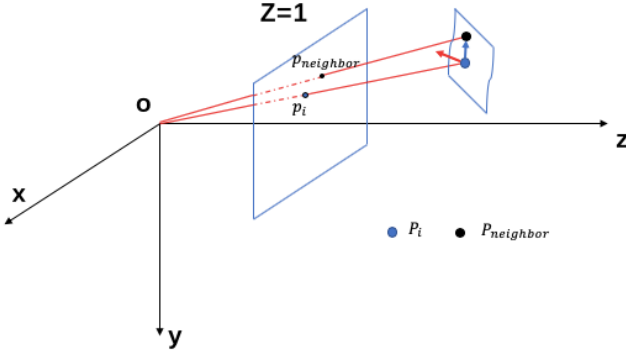$$N = \frac{1}{8}\sum_1^8 (N_i)$$

Fig. 4. Depth from the normal

When given the normal and initial depth, we can get the final refined depth. In Figure 4, for each pixel $p_i(x_i, y_i)$, the depth of neighbor point named $p_{neighbor}$ should be refined. Their corresponding 3D points are P_i and $P_{neighbor}$. Assuming that the normal of P_i is $N(N_x, N_y, N_z)$, the depth of P_i is Z_i, the depth of $P_{neighbor}$ is $Z_{neighbor}$, $N \perp P_i\vec{P}_{neighbor}$ is

Apparently reasonable due to the orthogonality and surface consistency in limiting field.

$$K^{-1}((x_i, y_i, 1)Z_i - (x_{neighbor}, y_{neighbor}, 1)Z'_{neighbor})\begin{bmatrix} N_x \\ N_y \\ N_z \end{bmatrix} = 0$$

For the refined depth, eight neighbor points are also adopted. Eight neighbor points are used to refined the depth of central point. Considering the discontinuity of normal in some edge or irregular surface of real object. The weight from the ref image is introduced to make depth more reasonable. The weight is defined as :

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

The W_{ij} depends on the gradient between p_i and $p_{neighbor}$, which means that the bigger gradient represents the less reliability of the refined depth. Considering the eight neighbor points, the final refined depth Z_{final} is a combination of weighted sum of eight different direction.

$$Z_{final} = \sum_1^8 w'_{ij}Z_{neighbor}$$

$$w'_{ij} == \frac{w_{ij}}{\sum_1^8 w_{ij}}$$

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

B. Pixel-Wise & Feature-Wise Loss Design

Due to the unsupervised method used here, how to design the loss function is more important. In this paper, 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 raise the completeness of point clouds.

The key points that are embodied in pixel-wise and feature-wise loss function is the consistency crossing multi views. Given the reference image I_{ref} and source image I_{src}, the corresponding intrinsicses are K_{ref} and K_{src}, the extrinsics from I_{ref} to I_{src} is T. For the pixel $p_i(x_i, y_i)$ in I_{ref}, the match pixel $p'_i(x'_i, y'_i)$ in I_{src} can be calculated as:

$$p'_i = K_{src}T(Z_{final}K_{ref}^{-1}p_i)$$

The overlapping area in source image I_{src} can be sampling using bilinear method, which can be called I'_{src}.

$$I'_{src} = I_{src}(p'_i)$$

For the occlusion area, the values in I'_{src} are set to zero. Obviously, the mask M can be obtained when the p_i is projected to the exterior of I_{src}.

Based on the constraint, the loss function L of NPF-MVSNet can be formulate as the equation.

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

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

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

Where m is the 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 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 the depth less steep in first order domain and second order domain.

$$L_{smooth} = \frac{1}{N} \sum (e^{-\alpha_2 |\bigtriangledown I_{ref}|} |\bigtriangledown Z| + e^{-\alpha_3 |\bigtriangledown^2 I_{ref}|} |\bigtriangledown^2 Z|)$$

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 pixel-wise loss, the main contribution of NPF-MVSNet is the use of feature-wise loss. For some textureless area, the pixel matching would be wrong, which lead to the low precision. But it will be changed with the aid of feature-wise loss. Using the more advanced information like semantic information, depth will be well learned even in the textureless regions.
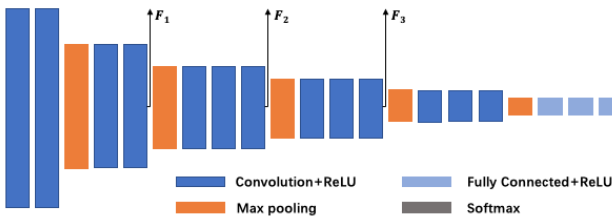

Fig. 5. feature-wise extraction from pre-trained VGG16

By putting the reference images I_{ref} into the pretrained 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 the size of the original input images. As a matter of fact, the layer 3 output the same size of the original input images, which is actually the performance 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 3.2., the matching pixel $p'_i(x'_i, y'_i)$ in F_{src} can be calculated. The matching feature map from F_{src} to F_{src} can be presented as below:

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

The estimated final depth will detect the similarity of feature maps. Beyond pixel value, the domain in feature have a bigger receptive field so that the textureless regions can be relieved in some extent. 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 feature maps.

$$L_{feature} = \beta_1 L_{F_1} + \beta_2 L_{F_2} + \beta_3 L_{F_3}$$

IV. Experiments

To prove the effective of our proposed NPF-MVSNet, this section mainly conducts lots of experiments. First we explore the performance of NPF-MVSNet on DTU dataset including the details of training and testing information. Then the current unsupervised networks in MVS are compared with NPF-MVSNet. In section 4.3., the ablation study is carried out to find the potential improvement in the network. At last, we use NPF-MVSNet on different dataset such as Tanks and Temples to study the generalization of our models.

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. By the lighting change, each scan has seven conditions with the pose known. We use the same train-validation-test split as in MVSNet [20] and MVS^2 [4]. 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: NPF-MVSNet 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 640×512. Due to the pyramid feature extraction, the resolution of final depth is 160×128. 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, with the learning rate of 1e-3 for the first 1 epochs and decreased 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 2 source images are used. In the process of testing, the resolution of input images is 1600×1200, which needs up to 10.612G GPU memory.

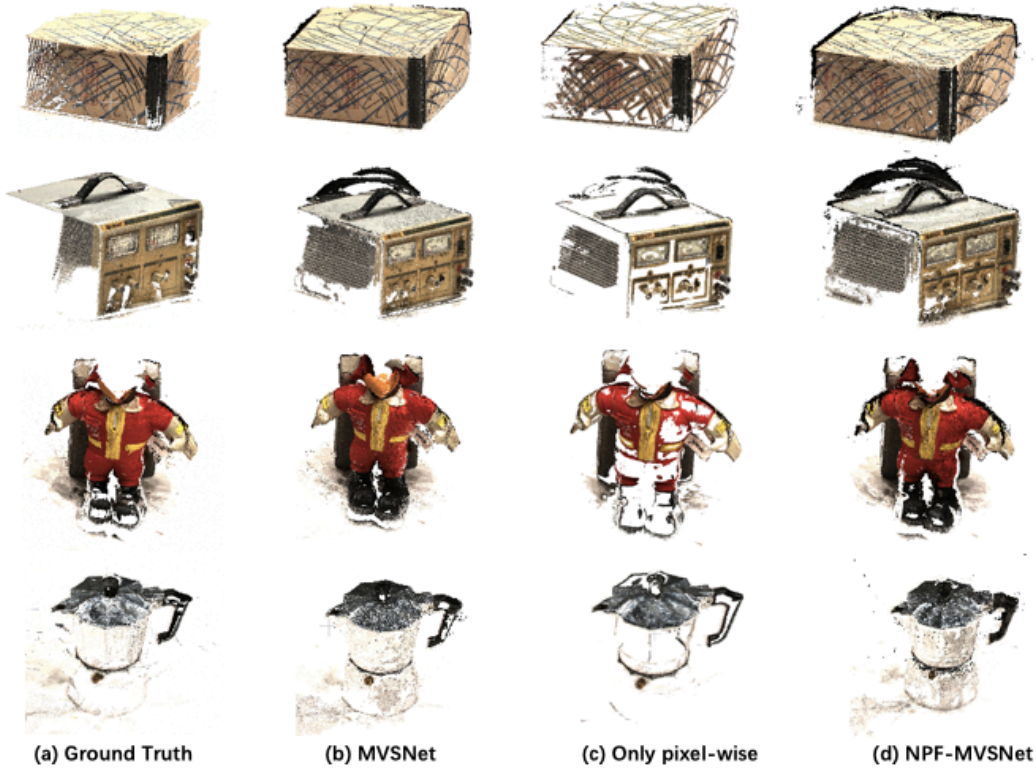(a) Ground Truth   (b) MVSNet   (c) Only pixel-wise   (d) NPF-MVSNet

Fig. 6. Qualitative comparison in 3D point clouds reconstruction between NPF-MVSNet and supervised or unsupervised MVS method on the DTU dataset. From left to right: ground truth, MVSNet(D=256) [20], the pixel-wise constraint, which is similar to unsup_mvs [11], and our proposed NPF-MVSNet.

2) Results on DTU: The official metric [8] is used to evaluate NPF-MVSNet' performance on DTU dataset. There are three metrics called accuracy, completion and overall.

To prove the effect of the model, we compare our proposed NPF-MVSNet against the three classic traditional methods such as Furu [5], Tola [18] and Colmap [17], and three classic supervised learning methods such as SurfaceNet [9], MVSNet [20] with different depth sample, and the only two unsupervised learning methods such as Unsup_MVS [11] and MVS^2 [4].

As table I shows, our proposed NPF-MVSNet can outperformance the two traditional methods and is so closed to Colmap [17]. As described in MVSNet [20], learning-based methods are ten times more efficient than traditional methods like Colmap. Further, due to the limitation of GPU memory, NPF-MVSNet select sampling value as 192 while the mean distances for MVSNet with different depth sampling value are given. Obviously, NPF-MVSNet surpasses the supervised learning method with the same setting. When compared with other unsupervised learning methods, conclusion can be made that our proposed NPF-MVSNet is the SOTA of unsupervised network for multi-view stereo reconstruction. For more details in point clouds, figure 6 illustrated the striking contrast. The reconstruction by NPF-MVSNet has more texture details. With the aid of feature-wise loss and pyramid feature, NPF-MVSNet can recover more textureless regions while normal-

Method	Mean Distance (mm)		
	Acc.	Comp.	overall
Furu [5]	0.612	0.939	0.775
Tola [18]	0.343	1.190	0.766
Colmap [17]	0.400	0.664	0.532
SurfaceNet [9]	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 [11]	0.881	1.073	0.977
MVS^2 [4]	0.760	0.515	0.637
NPF-MVSNet(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 BELOW.

depth consistency guarantees the accuracy of estimated depth in textureless area.

B. Comparison with Unsupervised Methods

There only two unsupervised methods until now. The first one is unsup_mvs [11], which is almost the first try in this direction. But it has poor performance while the overall of mean distance is 0.977. The other method published is MVS^2 [4]. Although it can reach to 0.637 in overall of mean distance,

it consumes around three times as much GPU memory than unsup_mvs, which is unaffordable for nvidia RTX 2080Ti used in NPF-MVSNet. To sum up, our propose unsupervised method achieves the best performance on the mean distance metric with the limited GPU memory.

C. Ablation Study

The section begins to analyze the effect of different modules proposed in NPF-MVSNet. There are mainly three contrast experiment carried out. We would explore the role of pyramid feature aggregation, normal-depth consistency and feature-wise loss design. The all experiments focus on the only one variable every time.

Pyramid Feature Aggregation. The module, which can catch more contextual information among different feature layer, is the enhanced version beyond the single feature map. Considering the limited GPU memory, we use the feature pyramid with only the finest level, which is the lite version in pyramid feature networks. As table II shows, this module will decrease the value of acc and comp in mean distance. To summarize, pyramid feature aggregation will improved 2% in overall.

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

Normal-depth consistency. From initial depth map to refined depth map, the module makes the depth map regularized in 3D space, which makes the depth more reasonable. Depth error is used to evaluated the quality of estimated depth before the reconstructed point clouds. Here we use percentage of predicted depths within 2mm, 4mm and 8mm of ground truth (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 objects 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

Figure 7 demonstrates the comparison with the case of only pixel-wise loss, with feature-wise loss but without normal-depth consistency, with both the feature-wise loss and normal-depth consistency. case (b) and case (c) are have better performance than case (a). But apparently, case (b) has more

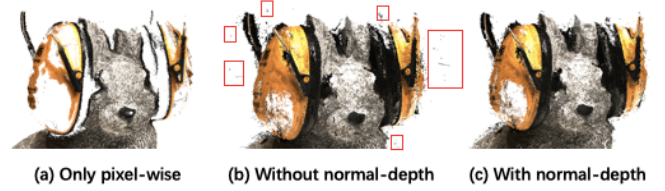(a) Only pixel-wise   (b) Without normal-depth   (c) With normal-depth

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

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 completeness. The explanation is that the 3D space regularization can guarantee the refined depth following the rule in real world. Figure 7 and table III can prove the benefits of this module.

Pixel-wise & Feature-wise Loss Design. The most unsupervised networks about depth estimation, either monocular video or binocular rectified image pairs, focus on the pixel-wise loss construction, which is Conform to humans' thoughts. But the constraint pointing to feature-wise is effective in previous related work [10] [19] [2]. We have compared the pixel-wise loss only and the addition of feature-wise loss, which shows the big improvement by the feature-wise component. What's more, how to select the the multi-scale feature maps is also taken into comparison.

As the telling in table IV, the overall of only pixel-wise loss is relatively higher. The addition of feature-wise loss makes the overall a impressive improvement. Further, we do ablation study on the different scale of feature from VGG16. The scale 1/4 is match to the resolution of depth by the network outputs. The results shows that the combination of 1/2, 1/4, 1/8 achieves the best results. By the way, adding the 1/8 feature maps improves the accuracy but deteriorate the completeness, which may be the cause of two higher semantic information is out of the control by 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 ON THE DIFFERENT LOSS DESIGN. WHERE THE SCALE 1/2 REPRESENTS THAT THE FEATURE (CORRESPONDING TO THE LAYER 8) EXTRACTED FROM THE PRE-TRAINED VGG16 NETWORKS IS HALF OF THE ORIGINAL INPUT IMAGES. THE SCALE 1/4, 1/8, 1/16 CORRESPONDS TO THE LAYER 15, 22, 29 IN VGG16

D. Generalization Ability on Tanks & Temples

To evaluate the performance of our unsupervised network on different dataset, the intermediate Tanks and Temples dataset, which has high resolution images of outdoor scenes of large

Method	Mean	Family	Francis	Horse	Lightouse	M60	Panther	Playground	Train
NPF-MVSnet	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 [12] AMONG UNSUPERVISED METHODS, WHICH IS FROM THE LEADERBOARD OF INTERMEDIATE T&T.

(a) Family   (b) Francis   (c)Horse   (d)M60

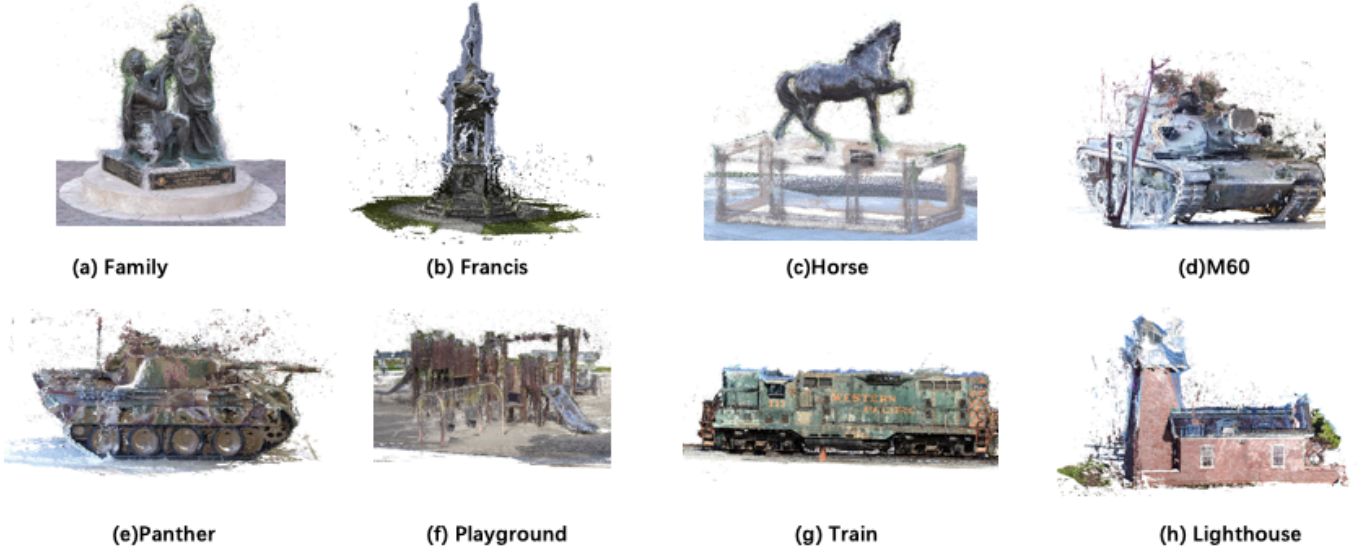(e)Panther   (f) Playground   (g) Train   (h) Lighthouse

Fig. 8. Our unsupervised network's performance on the Tanks and Temples dataset [12] without any finetuning.

objects, is adopted. The models of NPF-MVSNet 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 when the case of 192 depth intervals will out of memory. What's more, the another core hyperparameter is the photometric threshold in depth fusion. For the same depth maps of whole datasets, the different photometric threshold will cause the different performance threshold on the Tanks and Temples dataset. In other words, the hyperparameter will cause the change of accuracy and completeness. For NPF-MVSNet, the photometric threshold is set to 0.6 and we get the following results. As shows in table V, the ranking is selected from the leaderboard of intermediate T&T. NPF-MVSNet is better than MVS^2 by the mean score of 8 scenes, which is the best unsupervised MVS network until April 17, 2020. The reconstruction 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 NPF-MVSNet can be applied to advanced T&T but the reconstruction is so sparse due to the limitation of GPU memory. It's a game for 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 NPF-MVSNet, which achieve the state of arts in unsupervised MVS network. With our proposed methods of pyramid feature aggregation, normal-depth consistency and the feature-wise loss function along with pixel-wise loss function, NPF-MVSNet can capture contextual and semantic information from the perspective of perception, and make sure the rationality of estimated depth maps in real 3D world as to make it a best performance on DTU and other MVS datasets among all unsupervised network. In the future, more MVS datasets with high precision are desired. The domain transfer for different datasets can be improved better. Like prosperity in 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.

REFERENCES

[1] I. Alhashim and P. Wonka. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941*, 2018.
[2] W. Benzhang, F. Yiliu, F. Huini, and H. 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.

[3] N. D. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conference on Computer Vision*, pages 766–779. Springer, 2008.

[4] Y. Dai, Z. Zhu, Z. Rao, and B. Li. Mvs2: Deep unsupervised multi-view stereo with multi-view symmetry. In *2019 International Conference on 3D Vision (3DV)*, pages 1–8. IEEE, 2019.

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

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

[7] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. *arXiv preprint arXiv:1912.06378*, 2019.

[8] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413, 2014.

[9] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. 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.

[10] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[11] T. Khot, S. Agrawal, S. Tulsiani, C. Mertz, S. Lucey, and M. Hebert. Learning unsupervised multi-view stereopsis via robust photometric consistency. *arXiv preprint arXiv:1905.02706*, 2019.

[12] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

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

[14] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[15] K. Luo, T. Guan, L. Ju, H. Huang, and Y. 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.

[16] R. Mahjourian, M. Wicke, and A. 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.

[17] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys. Pixel-wise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016.

[18] E. Tola, C. Strecha, and P. Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23:903–920, 2011.

[19] A. Wang, Z. Fang, Y. Gao, X. Jiang, and S. Ma. Depth estimation of video sequences with perceptual losses. *IEEE Access*, 6:30536–30546, 2018.

[20] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[21] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. 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.

[22] H. Yi, Z. Wei, M. Ding, R. Zhang, Y. Chen, G. Wang, and Y.-W. Tai. Pyramid multi-view stereo net with self-adaptive view aggregation. *arXiv preprint arXiv:1912.03001*, 2019.

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