Detailed Facial Geometry Recovery from Multi-View Images by Learning an Implicit Function

Yunze Xiao[1*], Hao Zhu[1*], Haotian Yang[1], Zhengyu Diao[1], Xiangju Lu[2], Xun Cao[1†]

[1] Nanjing University, Nanjing, China
[2] iQIYI Inc, Beijing, China
{yunze.xiao, diaozhengyu}@smail.nju.edu.cn, {zhuhaoese, caoxun}@nju.edu.cn,
yanght321@gmail.com, luxiangju@qiyi.com

Abstract

Recovering detailed facial geometry from a set of calibrated multi-view images is valuable for its wide range of applications. Traditional multi-view stereo (MVS) methods adopt an optimization-based scheme to regularize the matching cost. Recently, learning-based methods integrate all these into an end-to-end neural network and show superiority of efficiency. In this paper, we propose a novel architecture to recover extremely detailed 3D faces within dozens of seconds. Unlike previous learning-based methods that regularize the cost volume via 3D CNN, we propose to learn an implicit function for regressing the matching cost. By fitting a 3D morphable model from multi-view images, the features of multiple images are extracted and aggregated in the mesh-attached UV space, which makes the implicit function more effective in recovering detailed facial shape. Our method outperforms SOTA learning-based MVS in accuracy by a large margin on the FaceScape dataset. The code and data are released in *https://github.com/zhuhao-nju/mvfr*.

Introduction

Recovering high-quality facial 3D models is of great value in many fields including the movie industry, facial analysis, and augmented reality. Multi-view stereo (MVS) focuses on reconstructing accurate shape from several calibrated images, which is the main way to obtain high-quality 3D face models. Traditional MVS pipeline mainly consists of matching cost computation, fusion, and refinement(Beeler et al. 2010; Furukawa and Ponce 2010; Galliani, Lasinger, and Schindler 2015), which can yield extremely fine facial geometry at the cost of tremendous calculations and running time. In recent years, learning-based MVS methods(Chen et al. 2019; Gu et al. 2020; Huang et al. 2018; Im et al. 2019; Ji et al. 2017; Kar, Häne, and Malik 2017; Luo et al. 2019; Yao et al. 2018, 2019) show superiority in accuracy and speed, which use a single-pass neural network instead of traditional iterative optimization. Though learning-based MVS outperforms traditional methods in certain scenes, there is still no learning-based MVS method that recovers fine-level facial geometry for industrial 3D modeling.

[*] These authors contributed equally to this work.
[†] Xun Cao is the corresponding author.

Figure 1: Our method reconstructs detailed facial geometry from multi-view images. The result shape is rendered in grey (left) and with normal-color (right). The input image is partially mosaicked for anonymization.

In this paper, we propose a novel framework to efficiently recover detailed facial geometry from calibrated multi-view images. The key observation is that the previous state-of-the-art learning-based methods(Yao et al. 2018, 2019; Chen et al. 2019; Gu et al. 2020; Yi et al. 2020; Yan et al. 2020) build and refine the cost volume in the camera frustum space, which is memory-consuming and limits the resolution. Though several works(Gu et al. 2020; Riegler, Ulusoy, and Geiger 2017; Wang et al. 2017) tried to improve the efficiency of the volume-based neural network, they didn't utilize the facial prior and fail to recover the detailed geometry. To overcome this problem, we propose to use an implicit function to regularize the matching cost to recover geometric details. To make the framework achievable, multi-view 3DMM estimation is utilized to transfer the cost space into UVD space (UV + displacement), which efficiently solves the rough shape, and supports the implicit function learning in detailed shape recovery. Compared to the previous learning-based MVS framework, our method doesn't require an additional fusion phase, which saves processing time and achieves higher accuracy.

The contribution can be summarized in three aspects:

- A novel facial-specific multi-view reconstruction pipeline is proposed consisting of base model fitting, learned implicit function, and mesoscopic prediction.
- We propose to tackle the problem of cost regression and multi-view fusion simultaneously by learning an implicit

function in our network, which is proved to be more accurate and time-saving.

- We propose to build the mesh-attached cost volume in UVD space, which greatly reduces the solution space for implicit function learning, achieving efficient and effective MVS for human faces.

Related Works

The related works can be divided into three categories: traditional MVS, learning-based MVS, and face-specific MVS.

Traditional MVS. Traditional MVS recovers the 3D shape of the objects or scenes from a set of calibrated multi-view images. According to the representation of the 3D modeling, traditional MVS can be categorized into volumetric method(Kutulakos and Seitz 2000; Seitz and Dyer 1997), point-based method(Furukawa and Ponce 2010; Lhuillier and Quan 2005; Chen et al. 2019) and depth-based method(Campbell et al. 2008; Galliani, Lasinger, and Schindler 2015; Schönberger et al. 2016; Tola, Strecha, and Fua 2012; Liu et al. 2009). Due to the space limitation, we recommend referring to the survey/benchmark(Seitz et al. 2006; Knapitsch et al. 2017; Zhu et al. 2017) and the tutorial(Furukawa and Hernández 2015) for the comprehensive review for traditional MVS. Here we focus on reviewing recently proposed learning-based MVS and face-specific MVS, which are more relevant to our work.

Learning-based MVS. Learning-based MVS methods seek to improve the reconstructing performance by introducing deep neural networks. In early attempts, Hartmann *et al.*(Hartmann et al. 2017) proposed to use the encoder network with multiple Siamese input branches to measure the similarity of the image patches from the multi-view input images. Ji *et al.*(Ji et al. 2017) proposed to use a CNN to predict each voxel a binary attribute indicating whether the voxel is on the surface. Kar *et al.*(Kar, Häne, and Malik 2017) proposed an end-to-end learning MVS network by involving differentiable feature projection and inverse-projection along viewing rays. Based on this framework, Yao *et al.* (Yao et al. 2018, 2019) built the cost volume on the camera frustum, and adopts 3D CNN or RNN to regularize the matching cost. Huang *et al.* (Huang et al. 2018) proposed to use the CNN to estimate the disparity from the plane-sweep volumes generated from multi-view images. Sunghoon *et al.*(Im et al. 2019) introduced the cost fusion and aggregation network, and finally regress a depth map from the refined cost volume. Chen *et al.*(Chen et al. 2019) proposed to process the target scene as point clouds in a neural network, which predicts the depth in a coarse-to-fine manner. Luo *et al.* (Luo et al. 2019) proposed the patch-based MVS network consisting of a patch-wise aggregation module to generate a matching confidence volume from extracted features, and a hybrid 3D CNN to infer a depth probability distribution. Gu *et al.*(Gu et al. 2020) extends the prior volume-based MVS network where cost volume is built upon a feature pyramid encoding geometry and context at gradually finer scales. Yan *et al.* (Yan et al. 2020) proposes a dense hybrid recurrent multi-view stereo net with dynamic consistency checking for accurate dense point cloud reconstruction. Yi *et al.*(Yi et al. 2020) presents a pyramid multi-view stereo network with the self-adaptive view aggregation. Yariv *et al.*(Yariv et al. 2020) proposes an implicit differentiable renderer that learns 3D geometry, appearance, and cameras from masked 2D images and noisy camera initialization. Different from all the above methods, our method is the first to learn an implicit function for cost regularization in the learning-based framework.

Face-specific MVS. Faces are special as they contain relatively fewer photometric features but plenty of geometric details. Based on the pyramid stereo matching framework, Beeler *et al.*(Beeler et al. 2010) uses smoothness constraint, ordering constraint, and uniqueness constraint to robustly recover the facial shape, then the shape is iteratively optimized to recover pore-level geometry. Bradley *et al.*(Bradley et al. 2010) extends the multi-view facial stereo to multi-view videos that are captured by an array of video cameras. Ghosh *et al.*(Ghosh et al. 2011) proposes to capture high resolution diffuse and specular photometric information using a multi-view face capture system, then reconstruct detailed facial geometry. Valgaerts *et al.*(Valgaerts et al. 2012) proposed a lightweight passive facial performance capture approach that only requires a single pair of stereo cameras. Though generalized MVS methods are able to recover fine 3D models, they consume a lot of computing resources and time to recover faithful geometric details. In recent years, many works study how to recover non-rigid facial geometry from uncalibrated images or videos. Dou *et al.*(Dou and Kakadiaris 2018) and Ramon *et al.*(Ramon, Escur, and Giro-i Nieto 2019) proposes to recover the face model from multi-view images using a subspace representation of the 3D facial shape and a deep recurrent neural network to fuse the identity-related features. Bai *et al.*(Bai et al. 2020, 2021) propose to optimize the 3D face shape by explicitly enforcing multi-view appearance consistency, which makes it possible to recover shape details according to conventional multi-view stereo methods.

Methods

As shown in Figure 2, our method consists of three stages. In the first stage, the 2D landmarks of multi-view images are extracted, and a base mesh is obtained by fitting the 3DMM to these landmarks. In the second stage, the accurate 3D facial geometry is predicted by learning the implicit function, and the detailed shape is refined by the mesoscopic prediction network in the last stage. We will explain each module in detail in the following sections.

Base Mesh Fitting

Firstly, we use Bulat *et al.*'s method (Bulat and Tzimiropoulos 2017) to obtain the 2D landmarks $[l_1, l_2, ..., l_M] \in \mathcal{R}^{M\times 2}$ for each image. Given the camera parameters, the 3D landmarks can be solved by minimizing the energy function below:

$$E(\mathcal{L}_j) = \sum_{i=0}^{N} ||\Pi_i(\mathcal{L}_j) - l_j||_2 \quad (1)$$

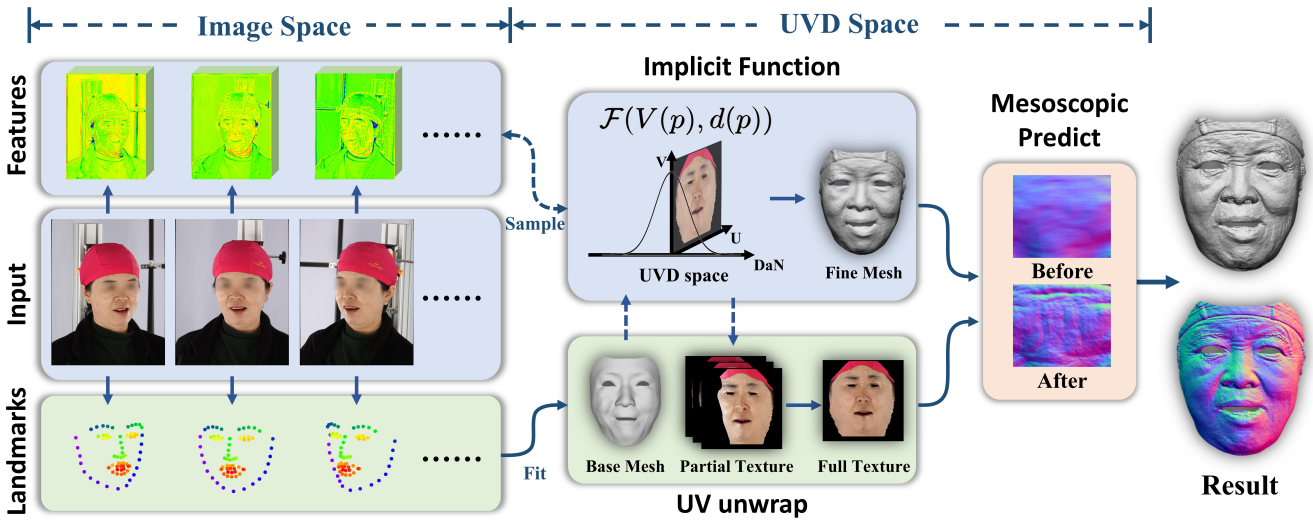

Figure 2: Our pipeline consists of three parts: base mesh fitting (green blocks), implicit function learning (blue blocks), and mesoscopic recovery (pink block).

where i is the index of the view, N is the total view number. Π_i is the perspective projection of the camera in ith view. $\mathcal{L}_j$ and l_j are the jth 3D and 2D landmarks separately. The $\mathcal{L}_j$ in $[\mathcal{L}_1, \mathcal{L}_2, ..., \mathcal{L}_M] \in \mathcal{R}^{M\times 3}$ is solved by least-squares respectively.

After 3D landmarks are obtained, we fit the bilinear model generated by FaceScape(Yang et al. 2020) to the 3D landmarks and obtain the topologically-uniformed base mesh. The bilinear model of FaceScape can be formulated as $V = \mathcal{B}(id, exp)$, where $\mathcal{B}$ is a linear mapping from the parameter id and exp to the vertices position V of the facial mesh model. The model fitting is solved by minimizing the energy function below:

$$E(\boldsymbol{id}, \boldsymbol{exp}, \mathcal{T}) = \sum_{j=0}^{M} ||\mathcal{T}(K_j(\mathcal{B}(\boldsymbol{id}, \boldsymbol{exp}))) - \mathcal{L}_j||_2 \quad (2)$$

where $\mathcal{T}()$ is the transformation between two coordinate systems. $K_j(\mathcal{M})$ extracts the jth 3D landmarks from the input mesh $\mathcal{M}$ using the predefined indices. We use the bilinear model generated by Yang *et al.*(Yang et al. 2020) in our experiments, and this can be substituted by other 3D facial parametric models.

Mesh-attached Cost Space

Different from the previous method that builds the cost volume in the Euclidean space or camera frustum, our method builds the cost volume in the UVD space attaching to the fitted facial mesh. The UV map is the flat 2D representation of the surface of the 3D mesh, where the attributes that are attached to the surface can be efficiently stored. We extend the UV space with the third dimension – displacement along the surface normal (DaN), and regularize the matching cost in this mesh-attached UVD space. This strategy is similar to the idea of SDF(Malladi, Sethian, and Vemuri 1995; Chan and Zhu 2005; Yariv et al. 2021) and has been used in many previous methods(Zhu et al. 2019, 2021b; Alldieck et al. 2019). We use the UV mapping defined by FaceScape, and the UV mapping is uniform in all fitted base meshes. The surface normal is obtained by sampling the vertex normal of the base mesh on the UV map, and the surface normal is smoothed to eliminate messy normal vectors caused by the high-frequency shape.

The establishment of mesh-attached volume brings three benefits to the facial MVS: 1) The mesh-attached volume connects the statistical facial model estimation and the volume-based depth regression method; 2) The solution space to represent geometry is reduced compared to the Euclidean volume(Kar, Häne, and Malik 2017) or camera frustum(Yao et al. 2018, 2019; Gu et al. 2020), thus the volume resolution can be enhanced; 3) The mesh-attached space can represent the complete face, while the camera frustum only sees the partial face and requires an additional fusion phase to produce complete facial geometry.

Implicit Function Learning

Implicit function. Inspired by the success of implicit function in single-view human shape recovery(Saito et al. 2019, 2020), we integrate implicit function learning into facial multi-view stereo. The overall network architecture is shown in Figure 3. Our implicit function $\mathcal{F}$ is formulated as:

$$\mathcal{F}(V(p), d(p)) = s, s \in \mathcal{R} \quad (3)$$

where $d(p)$ is the displacement corresponding to the sampled point p. $V(p)$ is the variance of the pixel-wise features extracted from multi-view images, formulated as:

$$V(p) = \frac{1}{N}\sum_{i=0}^{N} F_i^c(\Pi_i(p))^2 - (\frac{1}{N}\sum_{i=0}^{N} F_i^c(\Pi_i(p)))^2 \quad (4)$$

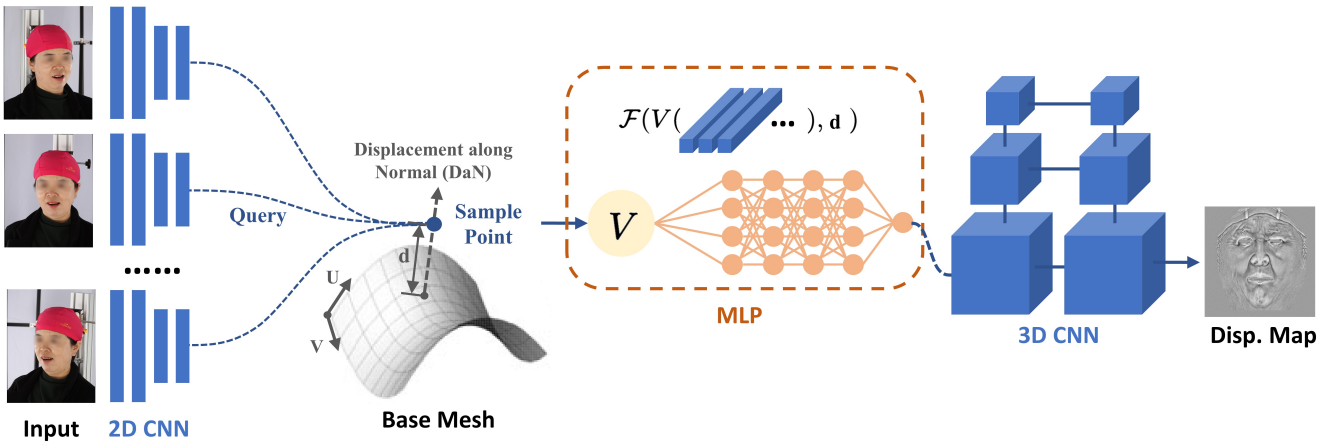

Figure 3: The network in implicit function learning stage. The weights-shared 2D CNN extracts the features of multi-view images. Then the points are sampled in UVD space, and the queried features are fed to a MLP to learn the implicit function. A 3D CNN based post-regularizer is used to extract the displacement map from the predicted cost volume.

where F_i^c is the feature with c channels extracted from ith image using the feature extracting network; Π_i is the projection function of the ith camera; N is the number of the input images.

We use the trainable multi-layer perceptron (MLP) to learn this implicit function. The label $\hat{s}$ for training is the distance-to-surface, which is the normalized probability to describe how close from the sampled point to the ground-truth surface. The label is formulated as:

$$\hat{s}(d) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(d-\hat{d})^2}{2{\sigma_1}^2}} \tag{5}$$

where σ_1 is set to 0.05 in our experiment. Our distance-to-surface label is different from the inside/outside label used in PIFu(Saito et al. 2019, 2020), and we will demonstrate the effectiveness in the ablation study section.

Feature extractor. We use a convolutional neural network to extract the features F_i^c in Equation 3 from the input images. The architecture of our feature extractor is modified based on the feature extractor of MVSNet(Yao et al. 2018) and RMVSNet(Yao et al. 2019). In our feature extractor, the group normalization(Wu and He 2018) is integrated into each convolutional layer, and half of the striding operations are removed to enhance the resolution of the generated feature maps. The feature extractor and the implicit function are trained in an end-to-end manner.

Surface-aware sampling. Compared to the CNN-based regularization network in the previous works(Yao et al. 2018, 2019; Gu et al. 2020; Yi et al. 2020; Yan et al. 2020), the implicit function benefits from the surface-aware sampling strategy, which uses more points close to the surface for training. This strategy has been used for human shape recovery in the 3D volumetric space(Saito et al. 2019, 2020), while we extend it to samples in the UVD space with a Gaussian distribution. Specifically, the probability distribution to sample in the UVD space follows Gaussian distribution in displacement dimension, and uniform distribution in U and V dimensions. The probability is formulated as:

$$P(u, v, d) = \frac{1}{\sqrt{2\pi}A} e^{-\frac{(d-\hat{d})^2}{2{\sigma_2}^2}} \tag{6}$$

where u, v, d are the coordinates along U, V, and displacement axes. $\hat{d}$ is the ground-truth displacement. A is the constant to normalize the integral value of probability to 1. In our experiments, we set $\sigma_2 = 0.1$ and the range of d is normalized to $[-1, 1]$. In order to sample the points far away from the surface, an additional uniform sampling in mesh-attached UVD space is adopted. We combine surface-aware sampled points and uniformly sampled points with a ratio of $16 : 1$.

Post-regularization. In the testing phase, the points in the complete UVD space are fed to the feature extractor and the implicit function, obtaining the cost volume in UVD space. A 3D convolutional neural network is used to further refine the cost volume, and the displacement map is extracted from the regularized cost volume using $softargmax$(Kendall et al. 2017) along the displacement dimension. Here we use $softargmax$ as it is the differentiable substitution of $argmax$. The predicted displacement map is applied to the base mesh, generating the final mesh.

Mesoscopic Prediction

The method in the last Section can recover fine middle-scale geometry, but still lacks the micro-scale geometry such as skin pores and minor wrinkles, because the disparity change across such a feature is too small to detect. Inspired by the traditional MVS method(Beeler et al. 2010), we notice the fact that light reflected by a diffuse surface is related to the integral of the incoming light, where the small concavities may block part of the incoming light and appear darker. This fact has been adopted in prior works(Beeler et al. 2010; Glencross et al. 2008) to synthesize the bump based on the

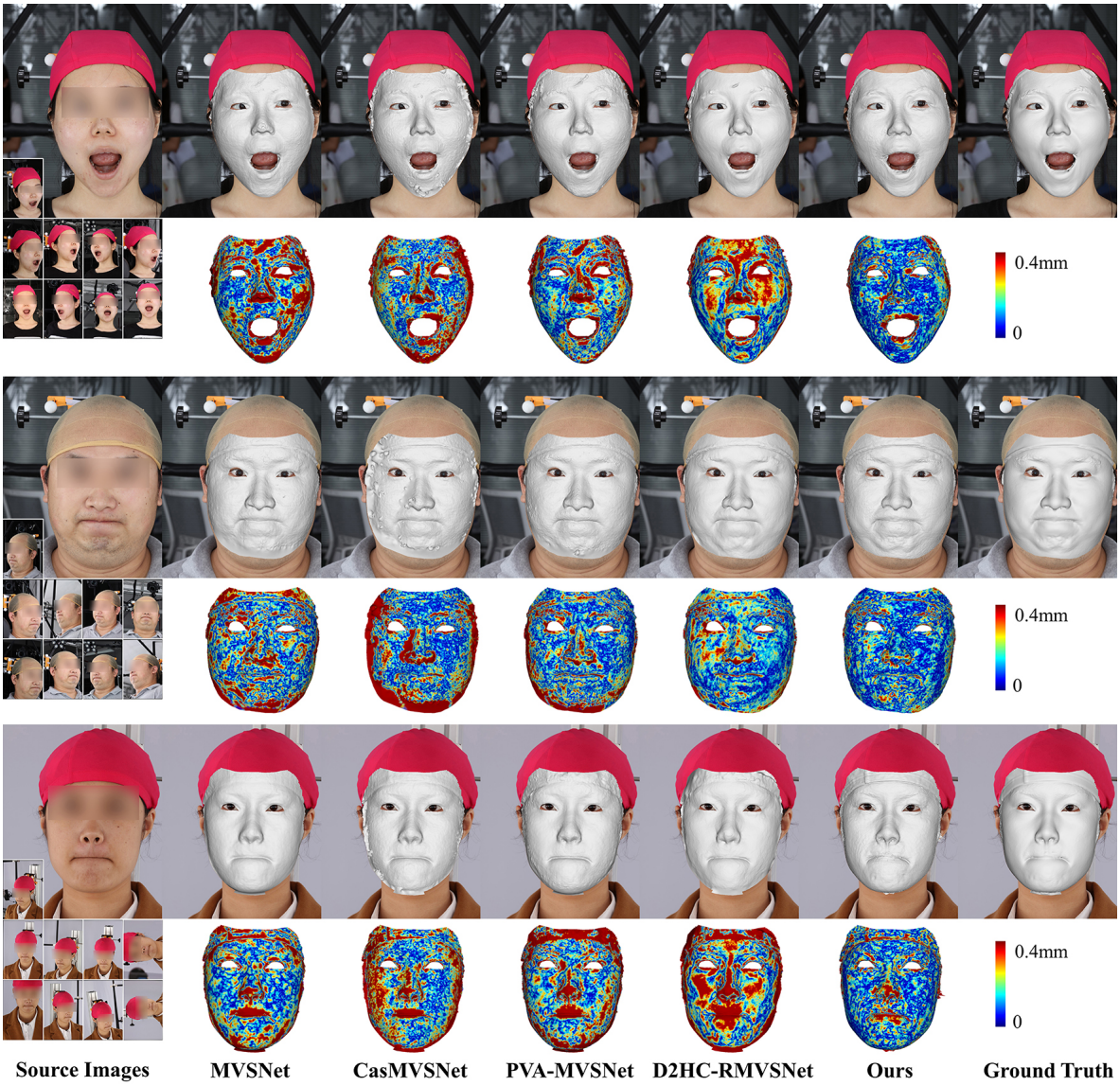

Figure 4: The rendered results and heat maps of ours and previous methods on FaceScape dataset.

rough mesh. Here we seek the potential to recover a displacement map representing the mesoscopic geometry from the multi-view images.

To be specific, we firstly use the resulting mesh to transfer the texture from the multiple images into the UV space, forming the partial UV textures. For each pixel in the UV texture, we compute the cosine of the angle between the normal direction and the camera direction as the weight, then blend these multi-view UV textures into a complete UV texture using the weighting average. The blended UV texture is fed to the mesoscopic network to predict the mesoscopic displacement map. The mesoscopic network uses the Pix2PixHD(Wang et al. 2018) network as the backbone and is trained using the displacement maps provided by FaceScape dataset.

Method	CD-mean/mm	CD-rms/mm	Comp./%
MVSNet	0.378 / 0.305	0.333 / 0.308	74.8 / 87.2
CasMVS	0.381 / 0.319	0.323 / 0.282	80.8 / 90.7
PVA	0.409 / 0.323	0.347 / 0.295	77.4 / 92.2
D2HC	0.382	0.324	83.5
Ours	**0.175**	**0.202**	**100.0**

* Numbers before and after '/' represent the original trained model and the fine-tuned model respectively.

Table 1: Comparison with Model-Generalized Methods

Experiments

Implementation Details

Data Preparation. We use FaceScape(Yang et al. 2020; Zhu et al. 2021a) dataset to train and validate our method. The

FaceScape dataset contains 7120 multi-view images and corresponding 3D models, each of which consists of roughly 50 images, corresponding calibrations, and the 3D mesh model. These data are captured from 359 different subjects and each in 20 expressions. We manually filter out the tuples with blur or calibration problems, then selected 80% of the remaining data as the training set and the other 20% as the testing set. For each tuple, we select 10 images from the frontal views as input. These images are scaled to 864×1296 for training and testing. The authorization of the data is described in the section of Ethics Statement.

In the base mesh fitting phase, we use the bilinear model generated by FaceScape dataset. The officially provided bilinear model is generated from 847 subjects, which contains the 359 subjects in our training/testing data. To fairly evaluate our methods, we regenerated the bilinear model with data that excluded the 359 subjects in our training/test set.

Training Details. The feature extractor and the implicit function are trained in an end-to-end manner, while the post-regularizer is trained after the other parameters are fixed. MSE loss is used to train the feature extractor + implicit function, and also the post-regularizer. We train our network using Adam optimizer, with the learning rate as 10^{-3}, and our model is trained in 200 epochs. The batch size is set to 1. We trained the model using Nvidia RTX 3090 for about 100 hours.

Metrics. We use chamfer distance (CD) between the predicted mesh and ground-truth mesh to measure the accuracy of the methods. We only measure the accuracy in the facial area, excluding eyes and inner mouth where the ground-truth geometry is hard to obtain. Considering the result meshes of generalized MVS methods(Yao et al. 2018; Gu et al. 2020; Yi et al. 2020; Yan et al. 2020) tend to extend out of the facial area, we mask out the outer mesh using the ground-truth mask of the frontal view. We run the experiments of comparison and ablation study on our testing sets, which contains 1341 tuples of data. For each tuple, there are images in 10 views, calibration parameters, and the ground-truth facial models. For all 1341 tuples of results, we report the mean of the chamfer distance (CD-mean) and the root mean square of chamfer distance (CD-rms). We also report the completeness (Comp.) for other methods, which is the area on the ground-truth mesh with error distance $< 2mm$ dividing the area of the full ground-truth mesh. When computing the chamfer distance, the area of the predicted mesh with error distance $> 2mm$ is filtered out. Considering our method recovers the complete face, the completeness of our method is labeled as 100%. All our mesh are counted when computing the chamfer distance.

Comparison with SOTA

We compare our method with previous methods quantitatively and qualitatively. The baselines are MVSNet(Yao et al. 2018), CasMVSNet(Gu et al. 2020), D2HC-RMVSNet(Yan et al. 2020), and PVA-MVSNET(Yi et al. 2020), which are state-of-the-art generalized MVS methods. We evaluate two models for these methods: 'ori' means the originally trained model provided by the authors, which were trained on DTU dataset(Aanæs et al. 2016; Jensen et al. 2014); 'ft' means the model fine-tuned using our training data generated from FaceScape dataset. The fine-tuned model is trained with epochs that are equivalent to our training settings. The scores of CD-mean, CD-rms, and completeness are reported in Table 1. The rendered results, as well as heat maps of error distance from predicted mesh to ground-truth mesh, are shown in Figure 4, and these results of previous methods are from the fine-tuned models.

From the quantitative comparison in Table 1, we can see that our method outperforms previous methods in CD-mean, CD-rms, and completeness for facial reconstruction. Among previous methods, though MVSNet(Yao et al. 2018) is slightly better in CD-mean and CD-rms than the other three methods, the completeness of MVSNet is worse, which means more bad areas are not counted in chamfer distance. We consider the reason is that our method benefits from the mesh-attached UVD cost space and the implicit function learning framework. The enhancement for each proposed module is discussed in the ablation study.

Besides, we compare our method with IDR(Yariv et al. 2020), which is a model-specific MVS method that also uses implicit learning for 3D reconstruction. IDR requires to be trained for each object, and the learned parameters are specific to the certain object. By contrast, our method and the methods mentioned above are model-generalized, which means they do not require to be trained for different input images. Since IDR takes an extremely long time to process (about 400 times slower than our method), we only compare the first 5 tuples of multi-view images, and the quantitative results are shown in Table 2. We can see that our method outperforms IDR by a large margin.

Data ID	1	2	3	4	5
IDR	0.971	0.744	1.129	1.065	1.003
Ours	0.173	0.130	0.149	0.195	0.181

Table 2: Comparison with Model-Specific Method.

Method	CD-mean	CD-rms
(a) Base	2.821	2.987
(b) Base+3DCNN	0.230	0.269
(c) Base+IF	0.195	0.212
(d) Base+IF(i/o)+Reg	0.181	0.222
(e) Base+IF+Reg	0.171	0.200
(f) Base+IF+Reg+Meso	0.175	0.202

Table 3: Ablation Study (unit:mm)

Ablation Study

We evaluate the performance with different components as defined below. For all the settings, the feature extractor and the base mesh fitting phase remain the same.

- (a) Base. The base mesh generated by bilinear model fitting.
- (b) Base+3DCNN. The implicit function prediction is replaced by 3D CNN regularizer. This strategy is adopted by prior works(Yao et al. 2018, 2019; Gu et al. 2020).

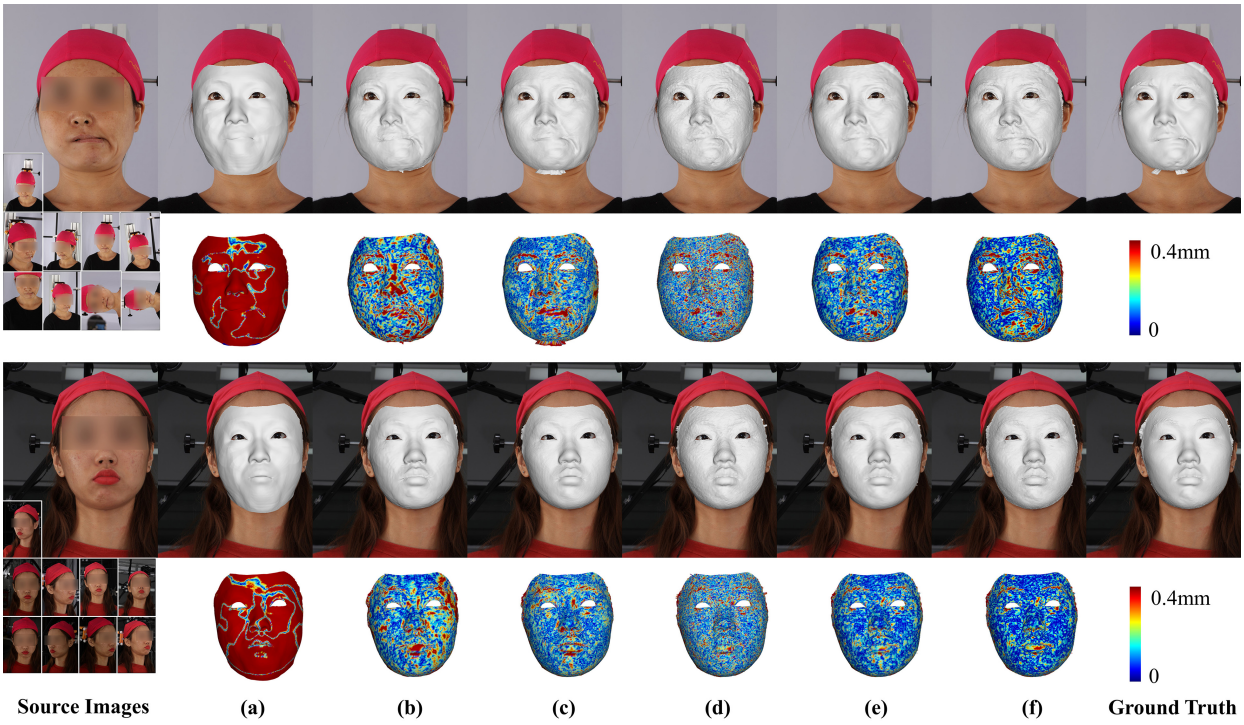

Figure 5: The rendered results and heat maps of the ablation study.

- (c) Base+IF. Post-regularize module and mesoscopic prediction are removed.
- (d) Base+IF(I/O)+Reg. Mesoscopic prediction is removed, and the implicit function here predicts the inside/outside labels, as used in (Saito et al. 2019).
- (e) Base+IF+Reg. Only mesoscopic prediction is removed.
- (f) Base+IF+Reg+Meso. Our complete method.

We visualize the results with different components in Figure 5, and report the quantitative evaluations of ablation study in Table 3. We can see that the base mesh in (a) is quite inaccurate, with chamfer distance $> 2mm$. Comparing (b) with (c), we can see that the implicit function-based framework is more accurate than 3D-CNN-based framework, and the performance of (c) can be further improved by adding a post-regularizer, as shown in (e). Comparing (d) and (e), we find that distance-to-surface labels are more effective than inside/outside labels. Comparing (f) with (e), though mesoscopic prediction doesn't enhance the accuracy quantitatively, it improves the visual effects by synthesizing mesoscopic geometry. Comparing (b) with MVSNet in Table 1, we can see that the cost volume in UVD space outperforms that in camera-frustum, since the much redundant space is ignored in the UVD space.

The performance for different view numbers are shown in Table 4. The mean chamfer distance from 10 views to 6 views fluctuates relatively little, but decrease severely when view points are fewer than 5. We believe that this is normal for the multi-view reconstruction method, because fewer viewpoints provide less matching information, resulting in unstable reconstruction.

View Num.	2	4	6	8	10
CD-mean	5.264	0.958	0.297	0.206	0.167
CD-rms	5.488	1.076	0.339	0.248	0.178

Table 4: Test on Different View Numbers (unit:mm)

Conclusion

In this paper, a novel architecture that leverages parametric model fitting and builds cost volume based on the attached UVD space to recover extremely detailed 3D faces. We propose to learn an implicit function for regularization in the task of multi-view stereo, which is proved to be more efficient and effective in recovering geometric details.

Limitation. There are still unresolved problems with our proposed method. Firstly, our method is only designed for rigid 3D facial reconstruction from well-calibrated and high-quality images. We found that the performance will degrade for the in-the-wild images, and the reason is that the resolution and signal-to-noise ratio of in-the-wild images are commonly too low for extremely fine 3D reconstruction. Secondly, our method can only recover a rigid 3D face, while cannot work for the image sequences with a dynamic face. An expression-identity disentangling module may be further added to solve this problem. Thirdly, it is hard to extend our method to arbitrary objects other than human faces, as our implicit function is built on the base of a parametric model.

Ethics Statement

The proposed method will promote the development of 3D facial modeling technology, which can be useful for applications requiring 3D face models with mesoscopic details, such as movies, games, and other entertainment industries. With the advancement of this technology, people may conveniently recover high-resolution and high-accuracy 3D face models from multi-view images through specialist devices and efficient algorithms. However, the efficient acquisition of detailed 3D facial models may cause severer privacy violation problems than 2D images. Therefore, we suggest that policymakers should establish an efficient monitoring platform to regulate the illegal spread of 3D face models with private information that may cause ethical problems.

Many images used in this paper are provided by FaceScape dataset(Zhu et al. 2021a; Yang et al. 2020), and the corresponding subjects have been known and authorized to use their photos for academic research. We have contacted the author of the FaceScape to confirm that the images in this article are authorized for publication. To protect the identity information of the subjects, we anonymized the face images by mosaicing the eyes. Please note that we used non-mosaic images for training and testing in the experiment, and only mosaic the images in the figures of this paper.

Acknowledgements

This work was supported by the NSFC grant 62025108, 62001213, and a gift funding from iQIYI Inc.

References

Aanæs, H.; Jensen, R. R.; Vogiatzis, G.; Tola, E.; and Dahl, A. B. 2016. Large-Scale Data for Multiple-View Stereopsis. *IJCV*, 1–16.

Alldieck, T.; Pons-Moll, G.; Theobalt, C.; and Magnor, M. 2019. Tex2shape: Detailed full human body geometry from a single image. In *ICCV*, 2293–2303.

Bai, Z.; Cui, Z.; Liu, X.; and Tan, P. 2021. Riggable 3D Face Reconstruction via In-Network Optimization. *arXiv preprint arXiv:2104.03493*.

Bai, Z.; Cui, Z.; Rahim, J. A.; Liu, X.; and Tan, P. 2020. Deep Facial Non-Rigid Multi-View Stereo. In *CVPR*, 5850–5860.

Beeler, T.; Bickel, B.; Beardsley, A. P.; Sumner, B.; and Gross, H. M. 2010. High-quality single-shot capture of facial geometry. *ToG*, 1–9.

Bradley, D.; Heidrich, W.; Popa, T.; and Sheffer, A. 2010. High resolution passive facial performance capture. In *SIGGRAPH*, 1–10.

Bulat, A.; and Tzimiropoulos, G. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230, 000 3D facial landmarks). *ICCV*.

Campbell, D. F. N.; Vogiatzis, G.; Hernández, C.; and Cipolla, R. 2008. Using Multiple Hypotheses to Improve Depth-Maps for Multi-View Stereo. *ECCV*, 766–779.

Chan, T.; and Zhu, W. 2005. Level set based shape prior segmentation. In *CVPR*, volume 2, 1164–1170.

Chen, R.; Han, S.; Xu, J.; and Su, H. 2019. Point-Based Multi-View Stereo Network. *ICCV*, 1538–1547.

Dou, P.; and Kakadiaris, I. A. 2018. Multi-view 3D face reconstruction with deep recurrent neural networks. *Image and Vision Computing*, 80: 80–91.

Furukawa, Y.; and Hernández, C. 2015. Multi-View Stereo: A Tutorial. *Foundations and Trends in Computer Graphics and Vision*, 1–148.

Furukawa, Y.; and Ponce, J. 2010. Accurate, dense, and robust multiview stereopsis. *PAMI*, 1362–1376.

Galliani, S.; Lasinger, K.; and Schindler, K. 2015. Massively Parallel Multiview Stereopsis by Surface Normal Diffusion. *ICCV*.

Ghosh, A.; Fyffe, G.; Tunwattanapong, B.; Busch, J.; Yu, X.; and Debevec, P. 2011. Multiview face capture using polarized spherical gradient illumination. In *SIGGRAPH Asia*, 1–10.

Glencross, M.; Ward, G. J.; Melendez, F.; Jay, C.; Liu, J.; and Hubbold, R. 2008. A perceptually validated model for surface depth hallucination. *ACM Transactions on Graphics (TOG)*, 27(3): 1–8.

Gu, X.; Fan, Z.; Zhu, S.; Dai, Z.; Tan, F.; and Tan, P. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *CVPR*, 2495–2504.

Hartmann, W.; Galliani, S.; Havlena, M.; Schindler, K.; and Gool, V. L. 2017. Learned multi-patch similarity. *ICCV*.

Huang, P.-H.; Matzen, K.; Kopf, J.; Ahuja, N.; and Huang, J.-B. 2018. DeepMVS: Learning Multi-view Stereopsis. *CVPR*, 2821–2830.

Im, S.; Jeon, H.-G.; Lin, S.; and Kweon, S. I. 2019. DPSNet: End-to-end Deep Plane Sweep Stereo. *ICLR*.

Jensen, R.; Dahl, A.; Vogiatzis, G.; Tola, E.; and Aanæs, H. 2014. Large scale multi-view stereopsis evaluation. In *CVPR*, 406–413.

Ji, M.; Gall, J.; Zheng, H.; Liu, Y.; and Fang, L. 2017. SurfaceNet: An End-to-end 3D Neural Network for Multiview Stereopsis. *ICCV*, 2326–2334.

Kar, A.; Häne, C.; and Malik, J. 2017. Learning a Multi-View Stereo Machine. *NIPS*, 364–375.

Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Kennedy, R.; Bachrach, A.; and Bry, A. 2017. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, 66–75.

Knapitsch, A.; Park, J.; Zhou, Q.-Y.; and Koltun, V. 2017. Tanks and temples: benchmarking large-scale scene reconstruction. *ToG*.

Kutulakos, N. K.; and Seitz, M. S. 2000. A Theory of Shape by Space Carving. *IJCV*, 199–218.

Lhuillier, M.; and Quan, L. 2005. A quasi-dense approach to surface reconstruction from uncalibrated images. *PAMI*, 418–433.

Liu, Y.; Cao, X.; Dai, Q.; and Xu, W. 2009. Continuous depth estimation for multi-view stereo. *CVPR*, 2121–2128.

Luo, K.; Guan, T.; Ju, L.; Huang, H.; and Luo, Y. 2019. P-MVSNet: Learning Patch-Wise Matching Confidence Aggregation for Multi-View Stereo. *ICCV*, 10452–10461.

Malladi, R.; Sethian, J.; and Vemuri, B. 1995. Shape Modeling with Front Propagation: A Level Set Approach. *PAMI*, 17(2): 158–175.

Ramon, E.; Escur, J.; and Giro-i Nieto, X. 2019. Multi-view 3D face reconstruction in the wild using siamese networks. In *ICCV Workshops*.

Riegler, G.; Ulusoy, O. A.; and Geiger, A. 2017. OctNet: Learning Deep 3D Representations at High Resolutions. *CVPR*.

Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Kanazawa, A.; and Li, H. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2304–2314.

Saito, S.; Simon, T.; Saragih, J.; and Joo, H. 2020. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 84–93.

Schönberger, L. J.; Zheng, E.; Frahm, J.-M.; and Pollefeys, M. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. *ECCV*, 501–518.

Seitz, M. S.; Curless, B.; Diebel, J.; Scharstein, D.; and Szeliski, R. 2006. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. *CVPR*, 519–528.

Seitz, M. S.; and Dyer, R. C. 1997. Photorealistic Scene Reconstruction by Voxel Coloring. *IJCV*, 151–173.

Tola, E.; Strecha, C.; and Fua, P. 2012. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Mach. Vis. Appl.*, 903–920.

Valgaerts, L.; Wu, C.; Bruhn, A.; Seidel, H.-P.; and Theobalt, C. 2012. Lightweight binocular facial performance capture under uncontrolled lighting. *ToG*, 31(6): 187–1.

Wang, P.-S.; Liu, Y.; Guo, Y.-X.; Sun, C.-Y.; and Tong, X. 2017. O-CNN: octree-based convolutional neural networks for 3D shape analysis. *ToG*.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *CVPR*.

Wu, Y.; and He, K. 2018. Group normalization. In *ECCV*, 3–19.

Yan, J.; Wei, Z.; Yi, H.; Ding, M.; Zhang, R.; Chen, Y.; Wang, G.; and Tai, Y.-W. 2020. Dense Hybrid Recurrent Multi-view Stereo Net with Dynamic Consistency Checking. In *ECCV*.

Yang, H.; Zhu, H.; Wang, Y.; Huang, M.; Shen, Q.; Yang, R.; and Cao, X. 2020. FaceScape: a Large-scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. In *CVPR*.

Yao, Y.; Luo, Z.; Li, S.; Fang, T.; and Quan, L. 2018. MVSNet: Depth Inference for Unstructured Multi-view Stereo. *ECCV*.

Yao, Y.; Luo, Z.; Li, S.; Shen, T.; Fang, T.; and Quan, L. 2019. Recurrent MVSNet for High-resolution Multi-view Stereo Depth Inference. *CVPR*.

Yariv, L.; Gu, J.; Kasten, Y.; and Lipman, Y. 2021. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34.

Yariv, L.; Kasten, Y.; Moran, D.; Galun, M.; Atzmon, M.; Ronen, B.; and Lipman, Y. 2020. Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance. *NIPS*, 33.

Yi, H.; Wei, Z.; Ding, M.; Zhang, R.; Chen, Y.; Wang, G.; and Tai, Y.-W. 2020. Pyramid multi-view stereo net with self-adaptive view aggregation. In *ECCV*.

Zhu, H.; Nie, Y.; Yue, T.; and Cao, X. 2017. The role of prior in image based 3d modeling: a survey. *Frontiers of Computer Science*, 11(2): 175–191.

Zhu, H.; Yang, H.; Guo, L.; Zhang, Y.; Wang, Y.; Huang, M.; Shen, Q.; Yang, R.; and Cao, X. 2021a. FaceScape: 3D Facial Dataset and Benchmark for Single-View 3D Face Reconstruction. *arXiv preprint arXiv:2111.01082*.

Zhu, H.; Zuo, X.; Wang, S.; Cao, X.; and Yang, R. 2019. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *CVPR*, 4491–4500.

Zhu, H.; Zuo, X.; Yang, H.; Wang, S.; Cao, X.; and Yang, R. 2021b. Detailed avatar recovery from single image. *PAMI*.