Learning Monocular Face Reconstruction using Multi-View Supervision

Zhixin Shu\textsuperscript{1,2}, Duygu Ceylan\textsuperscript{2}, Kalyan Sunkavalli\textsuperscript{2}
Eli Shechtman\textsuperscript{2}, Sunil Hadap\textsuperscript{2,3}, and Dimitris Samaras\textsuperscript{1}
\textsuperscript{1} Stony Brook University, Stony Brook, NY, USA
\textsuperscript{2} Adobe Research, San Jose, CA, USA
\textsuperscript{3} Amazon Lab126, Sunnyvale, CA, USA

Abstract—We present a method to reconstruct faces from a single portrait image. While traditional face reconstruction methods fit low-dimensional 3D morphable models to images, we train a deep network to regress depth from a single image directly. We do so by combining supervised losses on synthetic data with indirect supervision on real data using a novel multi-view photo-consistency loss. Furthermore, we regularize the depth estimation using a 3D morphable model (3DMM). We demonstrate that this leads to results that preserve facial features, capture facial geometry that goes beyond 3DMMs, and is also robust to viewpoint conditions. We evaluate our method on various datasets and via ablation studies, and demonstrate that it outperforms previous work significantly.

I. INTRODUCTION

Monocular 3D face reconstruction is an extensively studied problem with a wide variety of applications including face recognition [46], virtual avatar creation [51], image editing [3] and facial performance transfer [45], [24]. Traditional 3D face reconstruction methods represent facial geometry (and texture) using low-dimensional linear 3D morphable models (3DMM) [3], [8]. The 3DMM parameters can be estimated either by optimizing the geometric and photometric alignment of the 3DMM to the input image [3], [12], [45], [57] or by deep learning-based methods that regress 3DMM parameters from an image [33], [46], [44]. 3DMMs offer a low-dimensional and regularized shape space at the cost of limited expressive power. Thus, it is often not feasible to capture the geometric details that are important to a person’s identity.

As an alternative representation, other approaches estimate per-pixel depth [40] or a voxelized representation [19] from a single image. In principle, such representations are not limited to constrained shape space and given enough resolution can reconstruct details outside the space of 3DMMs. However, existing methods are trained with synthetic training data that is created using 3DMMs and thus inherently fail to capture the diversity of real faces.

In this work, our goal is to accurately reconstruct 3D geometry from a single image for faces “in-the-wild” with variations in viewpoint, lighting, and facial appearance. To this end, we propose to combine the flexibility and expressiveness of a depth-based representation with the regularized space of 3DMMs. We train a CNN that predicts per-pixel depth from the input image. Since depth estimation is an ambiguous task, we further incorporate a module that predicts 3DMM parameters from the image, which we use to regularize the depth estimation. At test time, we only use the depth estimation network to infer 3D face geometry that captures facial features that contribute to the identity of the person (see Figure 1).

Training such a network to handle real-world variations requires a large-scale training set of real images with per-pixel depth labels. Creating such a dataset is challenging and would require a complex, calibrated acquisition system. On the other hand, it is much easier to capture multiple images of a person (from multiple viewpoints), and this gives rise to a weaker form of supervision: two viewpoints of the same person when aligned using the correct depth should be photo consistent. We do not assume prior knowledge of the relative viewpoint transformation between the two images. Instead, we perform joint depth and viewpoint estimation in our network and use the resulting photometric error as a training loss. While similar photo-consistency losses have been employed for general depth estimation tasks [56], faces are non-Lambertian surfaces and are prone to self occlusions. These impose specific challenges that we address by introducing occlusion and suitability masks, which measure the

Fig. 1. We introduce a method for 3D face reconstruction from a single image using a depth-based representation. We obviate the need that real training images are paired with ground truth depth, pairing them instead with a second view of the same person and using multi-view supervision. Without post-processing refinement, our results capture the identity of the person better than a state-of-the-art 3DMM estimation method [43].
diffuseness of a pixel.

In principle, one could use multi-view stereo (MVS) methods [39] (that utilize a photo-consistency loss) to reconstruct geometry first, and use it to train a depth estimation network subsequently. However, MVS methods usually require a large number of calibrated images to produce a robust 3D reconstruction. In contrast, by directly embedding this as a loss function in our reconstruction network, we can learn depth estimation from a smaller set of uncalibrated images.

We motivate our network design via extensive ablation studies and show that we achieve state-of-the-art performance quantitatively. We also present qualitative results on challenging real test images. In summary, our contributions are the following:

- A novel network architecture that combines the flexibility of a depth map with the regularized shape space of a 3DMM to produce accurate and robust monocular 3D face reconstructions.
- A joint depth and viewpoint estimation network in conjunction with novel photometric losses (robust to non-Lambertian shading and self-occlusions) that allow training with real unlabeled images.
- Combining both synthetic and multi-view real image datasets (Multi-PIE [15] in our experiments), leading to state-of-the-art 3D face reconstruction performance both in quantitative and qualitative evaluations.

II. RELATED WORK

a) 3D Morphable Models: In their seminal work, Blanz and Vetter [3] introduced the notion of 3D Morphable Models (3DMMs) – a low dimensional linear representation for facial geometry and texture that they constructed by applying PCA to neutral face scans. In subsequent work, 3DMMs have been extended to include facial expressions [8] and significantly more facial variations [55] and non-linear morphable models [47]. Traditional 3D face reconstruction methods iteratively optimize the 3DMM parameters (pose and identity and expression coefficients) to best align the model to image cues including texture [4], image edges [35], facial landmark points [7], [12], [45], and optical flow [6]. Several methods also perform a joint fitting to multiple images of a subject [1], [30]. While such optimization methods often enable fast 3D face tracking, they either rely on statistical priors or good initialization and thus are prone to failures in challenging conditions. Moreover, they produce smooth reconstructions because of the 3DMM representation that are often refined using shape-from-shading approaches [11], [42], [37].

b) Learning 3D Face Reconstruction: Instead of relying on optimization-based methods, recent approaches have explored using deep neural networks to directly regress 3DMM parameters from a single image [21], [46], [57], [27], [44]. While such methods achieve impressive results, their main drawback is that they are limited to the shape space represented by the 3DMM. To overcome this limitation, alternative representations in the form of normal maps [41], depth maps [40], voxels [19], and dense facial correspondences [16], [54] have been proposed. Recent work [38] also explored training a generative model for unsupervised learning of 3D face shape. In our approach, we choose a depth map based representation which is flexible enough to capture geometric details. At the same time, we also incorporate a 3DMM estimation module to regularize the depth estimation.

Due to the difficulties in obtaining paired images and ground truth 3D shape information, many learning-based approaches rely on synthetic datasets created using 3DMMs [33], [34], [40]. While a synthetic data generation pipeline is ideal for generating all the necessary training signals, it is still a challenge to generalize models trained on such data to real images. In an attempt to address this challenge, Kim et al. [25] introduce a bootstrapping technique to match the distribution of the coefficients of the 3DMM sampled from the synthetic dataset to those of real images. As an alternative solution, several methods use an optimization-based method to fit 3DMMs to real images and use these fittings as ground truth for training [57], [46], [41]. Finally, more recent approaches have proposed unsupervised training strategies [43], [13], which incorporate a rendering module to render the predicted 3DMM and rely on an image-similarity loss. All these approaches commit to the 3DMM representation (either as the output representation or in the training data) and suffer from the loss of accuracy and generalizability that this approximation brings. In contrast, we use a more general depth map representation and train it with a photo-consistency loss on pairs of multi-view images of a subject captured at the same time instance. We show that this leads to more accurate and robust reconstructions.

c) Unsupervised Training for Depth Estimation: Several methods explore 3D inference from a single image by learning from registered 2D images, such as stereo pairs [10], [22], [14] or multi-view images with known camera parameters [52], [48]. Other methods assume having access to multiple images with no known relative viewpoint transformation [32], [49], [56]. Our approach is inspired by the recent success of these methods. However, faces are non-Lambertian surfaces, and the multi-view images often have large baselines that lead to self-occlusions. This leads to errors while applying photoconsistency losses based on image warping alone. We introduce novel loss functions to overcome these challenges.

III. APPROACH

A. Overview

Given a face image $I$, we seek to train an image-to-image translation network to estimate a depth image $D$, which encodes the per-pixel depth (i.e., distance from the camera) for the face region. We choose this representation because it can capture geometric details that are not represented by a 3DMM. We train this network with two types of data:

1. rendered synthetic data where we have access to paired image and depth maps which provide direct supervision, and
2. pairs of real images of the same subject from different viewpoints taken at the same time instance. While the latter type of data is crucial to generalize our method to unseen
real data, it does not provide a direct supervision signal. Instead, we utilize a photo-consistency loss by warping one of the given images to the other based on the predicted depth and comparing pixel similarity. We do not assume that the relative viewpoint transformation between the paired images is known, and instead estimate it with a view estimation sub-network. This allows us to generalize to uncalibrated capture setups where this information is not known apriori. Both to assist with viewpoint estimation and to also act as a regularizer for per-pixel depth estimation, we additionally incorporate a 3DMM prediction sub-network, which predicts the shape and expression parameters of a morphable model for each input image. Figure 2 shows the overall network architecture. During test time, we only utilize the depth estimation network. Next, we discuss different components of this architecture as well as the losses we employ for training.

B. Network Architecture

a) Depth Estimation Network (DepthNet): Given a face image \( I \in \mathbb{R}^{M \times N \times 3} \), we seek to train an image-to-image translation network, DepthNet, \( f(\cdot) \), to estimate a depth image \( D = f(I) \in \mathbb{R}^{M \times N \times 3} \). For each pixel \( I(i, j) \) in the face area, \( D(i, j) \) represents the distance of the pixel \( (i, j) \) from the camera. We use a U-Net architecture [36] for \( f \), similar to Sela et al. [40]. We refer to the supplementary material for details.

b) 3DMM Estimation Network (3DMMNet): Since depth prediction alone is an ambiguous task, we utilize 3DMMs for regularization. As in previous work [8], [11], [45], we represent the 3D face geometry, \( S \), as a 3D mesh with fixed topology and vertex positions computed as:

\[
S^{3D} = \bar{S} + \alpha_{id}G_{id} + \alpha_{exp}G_{exp},
\]

where \( \bar{S} \) denotes the average face shape and \( G = [G_{id}, G_{exp}] \) represent the identity and expression bases of the linear 3DMM; we use the bases from [29] and [8] for identity and expression respectively. Finally, \( \alpha = [\alpha_{id}, \alpha_{exp}] \) represent the image specific 3DMM identity and expression parameters respectively. This 3D mesh can be projected onto the 2D image as:

\[
S^{2D} = \Pi(RS^{3D} + t),
\]

where \( \Pi \) denotes the projection operator (in our experiments we employ perspective projection with fixed focal length) and \( R, t \) denote the extrinsic view transformation.

Given an input image \( I \), we train 3DMMNet, \( g(\cdot) \), that estimates both the 3DMM parameters, \( \alpha \), and the camera pose for \( I, V = (R, t) \). We use a DenseNet [17] architecture for \( g(\cdot) \) and provide details in the supplementary material.

We train the \( f(\cdot) \) and \( g(\cdot) \) networks in a hybrid manner using both synthetic data (where we have access to ground truth shape and 3DMM parameters) and real data (which consists of pairs of multi-view images which provide indirect supervision). Next, we describe the loss functions and datasets.

C. Learning 3DMM Inference: 3DMMNet

In this section, we detail the loss functions we use to train the 3DMM estimation network, \( g(\cdot) \).

a) 3DMM parameter regression.: Given an input image with ground truth 3DMM parameters \( \hat{\alpha} = [\hat{\alpha}_{id}, \hat{\alpha}_{exp}] \), we define an L1 loss to penalize the parameters, \( \alpha = [\alpha_{id}, \alpha_{exp}] \), predicted by \( g \) as:

\[
E^{\alpha}_{\alpha} = \lambda_1 ||\alpha_{id} - \hat{\alpha}_{id}|| + \lambda_2 ||\alpha_{exp} - \hat{\alpha}_{exp}||
\]

In case of synthetic data the ground truth parameters are known apriori. For real images, we use the identity shape regression network of Tran et al. [46] to predict \( \hat{\alpha}_{id} \) and treat this as ground truth. Since this network does not predict the expression parameters, we set \( \lambda_2 = 0 \) for real images. Otherwise we set \( \lambda_1 = \lambda_2 = 1.0 \).

b) Cross-image 3DMM parameter consistency.: When training our network with synchronized multi-view images of a subject, we enforce consistency in the inferred 3DMM parameters. In particular, this loss function enforces that inferred 3DMM parameters are consistent for pictures of...
the same person under different conditions such as varying viewpoints and pose. This does not require us to know the value of these parameters, but just their labels. As a result, we have the following consistency loss for image pair, \((I_1, I_2)\):

\[
E_{\text{pair}}^g = \|ga(I_1) - ga(I_2)\| + \|gexp(I_1) - gexp(I_2)\|. \tag{4}
\]

c) View estimation by landmarks re-projection.: As previously mentioned, given an input image, \(g\) predicts both the 3DMM parameters, \(\alpha\), and the camera pose \(\mathbf{V} = (\mathbf{R}, \mathbf{t})\). To learn camera pose, we use a loss based on 2D-3D correspondences. Specifically, we project a specific set of 3D landmark vertices from the 3DMM mesh, \(L^{3D}\), to the input image to obtain the projected 2D landmarks, \(L^{2D}\), and minimize their distance from the ground truth 2D landmark points detected on the input image, \(\hat{L}^{2D}\):

\[
E_{\text{land}}^g = \|L^{2D}(\alpha, \mathbf{V}) - \hat{L}^{2D}\| + \|\Pi(\mathbf{RL}^{3D} + \mathbf{t}) - \hat{L}^{2D}\|. \tag{5}
\]

Here \(L^{3D}\) depends on the 3DMM parameters, \(\alpha\). In case of synthetic data \(L^{3D}\) is known. For real images we use the network proposed in [5] to predict the 2D landmark points.

The combined training loss for \(g\) is then defined as:

\[
E^g = E^g_{\text{id}} + E^g_{\text{pair}} + E^g_{\text{land}}. \tag{6}
\]

All three losses are applied to both synthetic and real data, with the exception that when evaluating \(E^g_{\text{id}}\) for real data, we only consider the identity parameters.

D. Learning Depth Inference: DepthNet

When training the depth estimation network, we use a normalized depth representation to ease training. Specifically, we assume the tip of the nose is always at (0, 0, 0). We pre-crop the input images such that the tip of the nose is also roughly at the center of the image. Then, the \(z\)-coordinate of the camera location defines the distance between the face and the camera. Given a depth value \(D\), we convert it to \(D'\) by subtracting this camera-face distance (along the normal of the image plane) and normalizing the range to [0, 1]:

\[
D' = a(D - \xi_V) + c, \tag{7}
\]

where \(\xi_V\) denotes camera-face distance and is obtained from the view matrix \(V\). \(a\) and \(c\) are constant normalization factors to normalize the range of \(D'\) into [0, 1]. Both \(a\) and \(c\) are manually defined before training. In our experiments, we find that training with normalized depth values results in a more stable training process.

a) Supervision from synthetic data.: For synthetic data, we employ an \(L1\) loss between the predicted \(D = f(I)\) and ground truth \(\hat{D}\) depth values:

\[
E_{\text{depth}}^f = \|D - \hat{D}\|. \tag{8}
\]

In addition, following [40], we introduce a normal map regression loss which we find adds to the visual quality and the smoothness of the estimated depth map:

\[
E_{\text{normals}}^f = \|N_D - N_{\hat{D}}\|. \tag{9}
\]

\[N_D\] denotes the normal of each pixel in the image computed either from predicted or from ground truth depth values using finite differences.

b) Regularization with 3DMM.: Since the predicted 3DMM shape is expected to be accurate at a coarse scale, we use it to regularize the depth prediction. Specifically, given the predicted 3DMM parameters and the camera pose, we render a depth map from 3DMM to obtain \(D_S\). This rendering is done by projecting each vertex of the 3DMM into the image space and recording its depth value; we handle occlusions by z-buffering. We use \(D_S\) as a regularization term for the estimated depth using an \(L2\) loss function:

\[
E_{\text{multiview}}^f = \lambda_3\|D - D_S\|^2. \tag{10}
\]

We set \(\lambda_3 = 0.02\) and note that this loss is back-propagated only to DepthNet.

c) Multi-view photo-consistency.: With pairs of real/synthetic images of the same subject captured at the same time instance, \((I^1, I^2)\), we define a multi-view photo-consistency loss, as illustrated in Fig. 3, via view-synthesis, similar to SfM-Net [49]. Given the depth maps \((f(I^1), f(I^2))\) and the view parameters \((V^1, V^2)\) estimated for both images, we warp \(I^2\) to the viewpoint of \(I^1\) and measure the image similarity loss:

\[
E_{\text{multiview}}^f = \|I^1 - f(I^{2\rightarrow1})\|
= \|I^1 - w(I^2, V^1, V^2, f(I^2))\|, \tag{11}
\]

where \(I^{2\rightarrow1}\) represents the warped image and \(w\) represents the warping operation, which uses the estimated camera poses as well as the depth maps.

There are essential problems with applying such a view synthesis based loss for 3D face reconstruction. Variations in lighting, the non-Lambertian reflectance properties of the face, and self-occlusions caused by large differences in the two viewpoints naturally lead to violations of this constraint. Next, we describe how we make our photo-consistency loss robust to these problems.

Given a predicted shape \(S\) and two viewpoints \((V^1, V^2)\), not all parts of \(S\) are visible in both views. Therefore, we update the warping based image similarity loss to incorporate an occlusion mask:

\[
E_{\text{multiview}}^f = \|M^o \odot (I^1 - I^{2\rightarrow1})\|, \tag{12}
\]

Fig. 3. Multi-view photo-consistency loss function: For an image pair with known view parameters, a pixel in one view, corresponds to a 3D location and thus can be reprojected into the other view. Therefore the color differential for two corresponding pixels can be used to optimize the 3D location.
where $\odot$ represents the Hadamard product. The occlusion mask $M^o$ is a function of $(S, V^1, V^2)$:

$$M^o = m(S, V^1, V^2)$$  \hspace{1cm} (13)

Since we do not know the ground truth shape $S$, computing $M^o$ is not trivial. We approximate the computation of $M^o$ using predicted 3DMM parameters. To simplify the training process, we warp the morphable model depth from $V^1$ to $V^2$ using the bilinear sampler and compute the discrepancy with the original 3DMM depth at $V^2$. We apply a threshold (0.05 in our experiments) on the discrepancy to decide if this pixel is occluded or not.

An inherent problem in measuring the photo-consistency loss for structure-from-motion (or image warping based losses) is that the same physical point may lead to inconsistent pixel colors across different views. This multi-view inconsistency is usually due to varying illumination conditions and non-Lambertian surfaces (see Figure 4). One possible solution is to identify such inconsistencies in a pre-processing stage. For example, S2Dnet [50] uses synthetic data to train a network to remove the specular component of surface appearance and produces diffuse-only images that can be used in a multi-view stereo method. However, generating synthetic face images with realistic reflectance properties (e.g., sweat, makeup, oily skin) is not a trivial task.

Therefore, we opt for the alternative approach of training a suitability network, $h(\cdot)$, which predicts a “suitability map” $C = h(I) \in \mathbb{R}^{N \times N \times 1}$ for an image $I$ where each pixel $C(i,j)$ encodes how suitable the corresponding pixel is for a photo-consistency loss, e.g., a measure of its diffuseness. Instead of enforcing this map to be binary, we use a continuous weighting scheme and update the view consistency loss to be:

$$E^f_{VS} = \| h(I^1) \odot M^o \odot (I^1 - I^{(2\rightarrow 1)}) \|,$$  \hspace{1cm} (14)

where $\odot$ represents the Hadamard product. We train $h$ in an unsupervised way and to avoid the trivial solution of all zeros, we define a regularization loss:

$$E^h_{reg} = \| C - I \|,$$  \hspace{1cm} (15)

Note that our suitability map estimation is similar to the work of Zhou et al. [56], which also predicts an explainability map between two views to model occlusions and non-diffuse surfaces. Similar ideas have also been applied to facial expression editing [31], where an editing mask is produced, in an unsupervised manner, for the purpose of localized edits. Since we have an intermediate shape representation via the 3DMM inference, in our work, we decouple inconsistencies due to occlusions and non-diffuse surfaces. While we approximate occlusions analytically with the 3DMM shape, we use an additional network only to estimate suitability based on surface reflectance properties.

In summary, the final objective function for training is:

$$E^{f,h} = E^{f}_{depth} + E^{f}_{normals} + E^{f}_{3DMM} + E^{f,h}_{multiview} + E^{h}_{reg}.$$  \hspace{1cm} (16)
start training DepthNet along with the suitability network while keeping the parameters of 3DMMNet fixed. In both cases, we use the Adam [26] optimizer with a learning rate of 0.0002. At test time, we only use DepthNet for inference.

V. Evaluation

A. Qualitative Evaluation

We provide qualitative results of our method in Figures 1 and 6 by testing on random real images. More results are provided in the supplementary material. In addition, we provide qualitative comparisons to state-of-the-art methods that utilize different 3D representations. Figure 7 shows comparisons to the method of Sela et al. [40] which also uses a depth-map based representation. With both synthetic and real images in training, our method generalizes to unseen real data better and predicts much more accurate depth maps. In Figure 8, we show comparisons to state-of-the-art methods that utilize a voxel [19] and 3DMM based representations [43]. Although these methods provide smooth reconstructions, our results better preserve the identity of the subject by capturing facial features outside the shape space of a 3DMM or a limited voxelized volume.

We note that DepthNet implicitly learns to predict depth for only the face region without any special treatment of the background. This is possible because of two factors. First, for synthetic data, the ground truth depth is defined only for the face region. Second, through 3DMM regularization, we are enforcing the network to predict a depth value of 0 outside the face region. Our method runs at 50fps on an NVIDIA TitanX GPU, and both the input image and predicted depth map are at resolution 360 × 360.

B. Quantitative Evaluation

We evaluate our method on the BU-3DFE dataset [53] and renderings of 3D faces from the Florence dataset [2]. BU-3DFE provides images and ground truth scans of 100 subjects. Following the evaluation protocol of [40], we compare the output of DepthNet to this ground truth shape. We perform this comparison within the face region where we have per-pixel depth estimates. Since the coordinate systems used by our network and the BU-3DFE dataset are different,
we first perform a dataset-wise average global similarity transform (global scaling and translation) using the positions of 9 landmark points. We then perform ICP between our predicted depth map and the ground truth 3D shape to compute closest point correspondences. We measure the distance between such correspondences to compute the reconstruction accuracy. We use ICP only to compute correspondences, the rigid alignment is not applied to our predicted depth maps. We compare our results with previous methods in Table I, where (for normalization purposes) the numbers represent percentiles of the ground-truth depth range. We also provide further analysis of the reconstruction accuracy for different facial expressions in the dataset (Table II). Our method achieves superior performance and has consistent accuracy across different expressions.

Table I

Quantitative evaluation on BU-3DFE.

<table>
<thead>
<tr>
<th>Mean Err.</th>
<th>Std Err.</th>
<th>Median Err.</th>
<th>90% Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[23]</td>
<td>3.89</td>
<td>4.14</td>
<td>7.34</td>
</tr>
<tr>
<td>[58]</td>
<td>3.85</td>
<td>3.23</td>
<td>7.91</td>
</tr>
<tr>
<td>[33]</td>
<td>3.61</td>
<td>2.99</td>
<td>6.82</td>
</tr>
<tr>
<td>[40]</td>
<td>3.51</td>
<td>2.69</td>
<td>6.59</td>
</tr>
<tr>
<td>Ours</td>
<td>2.95</td>
<td>2.83</td>
<td>3.26</td>
</tr>
</tbody>
</table>

Table II

The mean error for different expressions in BU-3DFE. Left to right: Anger, Disgust, Fear, Happy, Neutral, Sad, Surprise.

<table>
<thead>
<tr>
<th>AN</th>
<th>DI</th>
<th>PE</th>
<th>HA</th>
<th>NR</th>
<th>SA</th>
<th>SU</th>
</tr>
</thead>
<tbody>
<tr>
<td>[23]</td>
<td>3.47</td>
<td>4.03</td>
<td>2.94</td>
<td>4.30</td>
<td>3.45</td>
<td>3.52</td>
</tr>
<tr>
<td>[58]</td>
<td>4.00</td>
<td>3.93</td>
<td>3.91</td>
<td>3.70</td>
<td>3.76</td>
<td>3.61</td>
</tr>
<tr>
<td>[33]</td>
<td>3.42</td>
<td>3.46</td>
<td>3.64</td>
<td>3.41</td>
<td>4.22</td>
<td>3.39</td>
</tr>
<tr>
<td>[40]</td>
<td>3.67</td>
<td>3.34</td>
<td>3.36</td>
<td>3.01</td>
<td>3.17</td>
<td>3.37</td>
</tr>
<tr>
<td>Ours</td>
<td>2.96</td>
<td>3.00</td>
<td>2.85</td>
<td>2.77</td>
<td>2.67</td>
<td>2.95</td>
</tr>
</tbody>
</table>

Fig. 10. We compare (a) training DepthNet only using our synthetic dataset (without Multi-PIE) and (b) training with both synthetic and real images. Training with real images enables the model to better generalize to real testing images (notice the misalignment between shape and texture in (a)–first row.)
as misalignment between shape and texture. Please see the supplementary material for more ablation experiments.

Table III

NME on 3D face renderings from the Florence dataset.

<table>
<thead>
<tr>
<th></th>
<th>Ours</th>
<th>[19]</th>
<th>[57]</th>
<th>[18]</th>
<th>[9]</th>
</tr>
</thead>
<tbody>
<tr>
<td>NME</td>
<td>0.0475</td>
<td>0.0509</td>
<td>0.0975</td>
<td>0.1253</td>
<td>0.0362</td>
</tr>
</tbody>
</table>

VI. CONCLUSIONS AND FUTURE WORK

We presented a single-image based 3D face reconstruction network using a depth-map based representation flexible enough to capture important geometric features for human faces. We also employed a 3DMM module for regularization. We trained our network by supervised losses on synthetic data and indirect supervision on multi-view real images using a photo-consistency loss. Training on real images made our network more robust to variations of viewpoint, illumination, and expression. Both qualitative and quantitative results demonstrated the benefit of our approach.

We used the Multi-PIE [15] dataset to illustrate the effectiveness of a multi-view consistency loss. Our method, however, is not restricted to images in lab settings since we do not assume any calibration information to be provided. Working with multi-view images captured in the wild is an exciting future direction. While we focus on geometry estimation, using post-processing methods to refine the geometry further and estimate appearance is also possible.

VII. ACKNOWLEDGEMENTS

This work was supported in part by a gift from Adobe, NSF grants CNS-1718014 and IIS-1763981, NIH grant NICHD 1R21 HD93912-01A1, the Partner University Fund, and the SUNY2020 Infrastructure Transportation Security Center. We would also like to thank Haoxiang Li for his valuable comments.

REFERENCES
