Learning Monocular Face Reconstruction using Multi-View Supervision

Zhixin Shu[1,2], Duygu Ceylan[2], Kalyan Sunkavalli[2]
Eli Shechtman[2], Sunil Hadap[2,3], and Dimitris Samaras[1]
[1] Stony Brook University, Stony Brook, NY, USA
[2] Adobe Research, San Jose, CA, USA
[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).

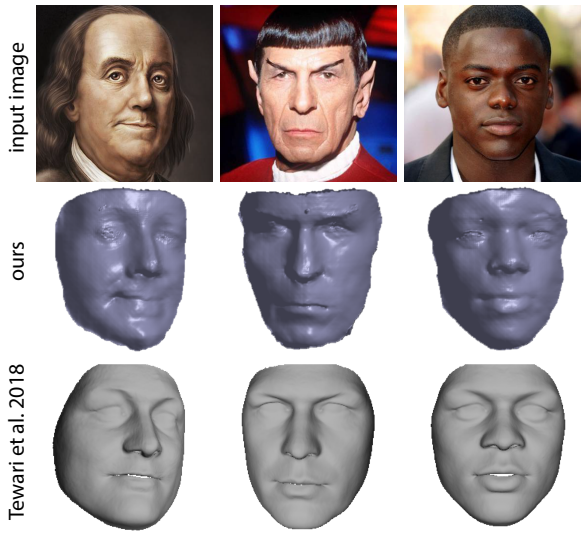

Fig. 1. We introduce a method for 3D face reconstruction from a single image using a depth-based representation. We obviate the need that real training images are paired with ground truth depth, pairing them instead with a second view of the same person and using multi-view supervision. Without post-processing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

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

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

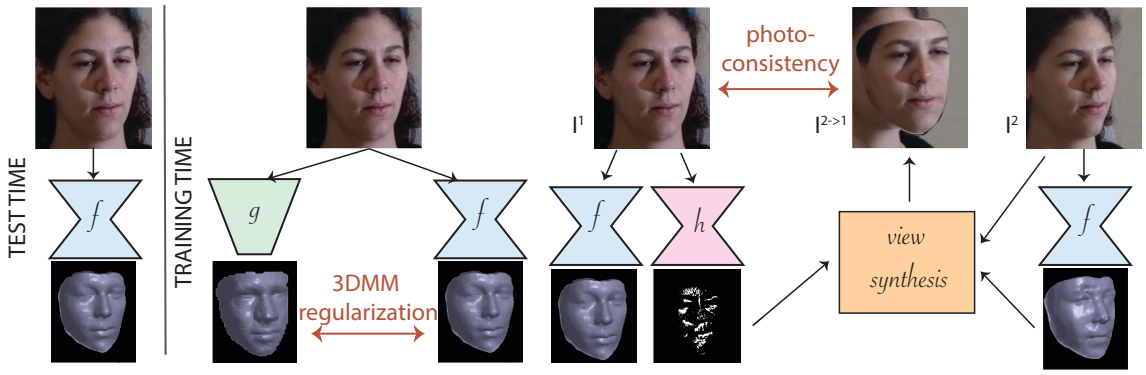

Fig. 2. Our goal is to train DepthNet, $f(\cdot)$, which predicts a depth map from a single input image at test time. During training, we also train 3DMMNet, $g(\cdot)$, which predicts 3DMM parameters from an input image. We define a loss that forces the depth predicted by $f(\cdot)$ to be similar to the depth obtained from the 3DMM. We also train our network with pairs of multi-view images (I^1, I^2) of the same subject. We use the predicted depth maps for these images to warp one view to the other ($I^{2\rightarrow1}$) and define a photo-consistency loss. This loss is weighted by a *suitability mask* predicted by $h(\cdot)$. This mask measures the diffuseness of pixels in the input and avoids computing the photo-consistency error in regions with self-occlusions and non-Lambertian surface properties.

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 $I \in R^{M\times N\times 3}$, we seek to train an image-to-image translation network, *DepthNet*, $f(\cdot)$, to estimate a depth image $D = f(I) \in R^{M\times N\times 1}$. 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:

$$S^{3D} = \bar{S} + \alpha_{id}G_{id} + \alpha_{exp}G_{exp}, \tag{1}$$

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:

$$S^{2D} = \Pi(\mathbf{R}S^{3D} + \mathbf{t}), \tag{2}$$

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

Given an input image I, we train *3DMMNet*, $g(\cdot)$, that estimates both the 3DMM parameters, α, and the camera pose for I, $\mathbf{V} = (\mathbf{R}, \mathbf{t})$. 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 $\hat{\alpha} = [\hat{\alpha}_{\text{id}}, \hat{\alpha}_{\text{exp}}]$, we define an $L1$ loss to penalize the parameters, $\alpha = [\alpha_{\text{id}}, \alpha_{\text{exp}}]$, predicted by g as:

$$E_\alpha^g = \lambda_1\|\alpha_{\text{id}} - \hat{\alpha}_{\text{id}}\| + \lambda_2\|\alpha_{\text{exp}} - \hat{\alpha}_{\text{exp}}\| \tag{3}$$

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 $\hat{\alpha}_{\text{id}}$ 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):

$$E^g_{\text{pair}} = \|g_{\text{id}}(I_1) - g_{\text{id}}(I_2)\| + \|g_{\text{exp}}(I_1) - g_{\text{exp}}(I_2)\|. \quad (4)$$

c) View estimation by landmarks re-projection.: As previously mentioned, given an input image, g predicts both the 3DMM parameters, α, 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, $\hat{L}^{2D}$:

$$E^g_{\text{land}} = \|L^{2D}(\alpha, \mathbf{V}) - \hat{L}^{2D}\| = \|\Pi(\mathbf{R}L^{3D} + \mathbf{t}) - \hat{L}^{2D}\|. \quad (5)$$

Here L^{3D} depends on the 3DMM parameters, α. In case of synthetic data $\hat{L}^{2D}$ 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:

$$E^g = E^g_\alpha + E^g_{\text{pair}} + E^g_{\text{land}}. \quad (6)$$

All three losses are applied to both synthetic and real data, with the exception that when evaluating E^g_α 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 $[0, 1]$:

$$D' = a(D - \xi_V) + c, \quad (7)$$

where ξ_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:

$$E^f_{\text{depth}} = \|D - \hat{D}\|. \quad (8)$$

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:

$$E^f_{\text{normals}} = \|N_D - N_{\hat{D}}\|. \quad (9)$$

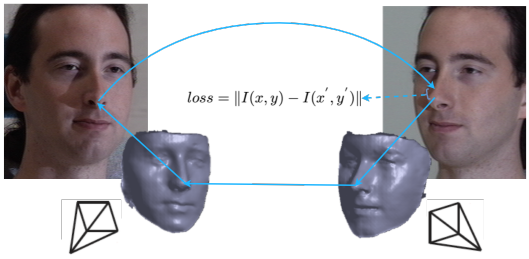

Fig. 3. Multi-view photo-consistency loss function: For an image pair with known view parameters, a pixel in one view, corresponds to a 3D location and thus can be reprojected into the other view. Therefore the color differential for two corresponding pixels can be used to optimize the 3D location.

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:

$$E^f_{\text{3DMM}} = \lambda_3 \|D - D_S\|^2. \quad (10)$$

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 $(\mathbf{V}^1, \mathbf{V}^2)$ estimated for both images, we warp I^2 to the viewpoint of I^1 and measure the image similarity loss:

$$\begin{aligned} E^f_{\text{multiview}} &= \|I^1 - I^{(2\to1)}\| \\ &= \|I^1 - w(I^2, \mathbf{V}^1, \mathbf{V}^2, f(I^2))\|, \end{aligned} \quad (11)$$

where $I^{(2\to1)}$ 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 $(\mathbf{V}^1, \mathbf{V}^2)$, not all parts of S are visible in both views. Therefore, we update the warping based image similarity loss to incorporate an occlusion mask:

$$E^f_{\text{multiview}} = \|M^o \odot (I^1 - I^{(2\to1)})\|, \quad (12)$$

where $\odot$ represents the Hadamard product. The occlusion mask M^{o} is a function of $(S, \mathbf{V}^1, \mathbf{V}^2)$:

$$M^{\text{o}} = m(S, \mathbf{V}^1, \mathbf{V}^2) \tag{13}$$

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 $\mathbf{V}^1$ to $\mathbf{V}^2$ using the bilinear sampler and compute the discrepancy with the original 3DMM depth at $\mathbf{V}^2$. 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" $C = h(I) \in R^{M \times N \times 1}$ 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:

$$E_{\text{VS}}^{f,h} = \|h(I^1) \odot M^{\text{o}} \odot (I^1 - I^{(2\to1)})\|, \tag{14}$$

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:

$$E_{\text{reg}}^{h} = \|C - 1\|, \tag{15}$$

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:

$$E^{f,h} = E_{\text{depth}}^{f} + E_{\text{normals}}^{f} + E_{\text{3DMM}}^{f} + E_{\text{multiview}}^{f,h} + E_{\text{reg}}^{h}. \tag{16}$$

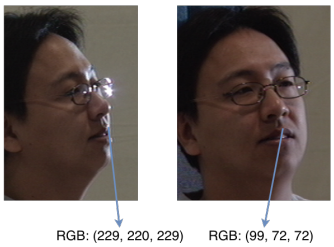

Fig. 4. The assumption of color-consistency is often violated because of different camera sensors or surface reflectance, as shown in this example of two synchronized Multi-PIE images from different views.

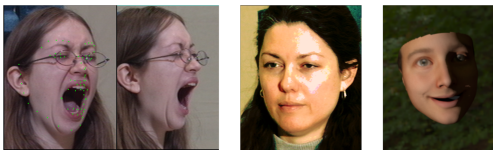

Fig. 5. For training, we utilize paired image from Multi-PIE with pre-computed landmark annotations (left), color augmented Multi-PIE images (middle), and synthetic faces (right).

While E_{depth}^{f} and E_{normals}^{f} are computed for only synthetic data, the remaining losses are computed both for real and synthetic data.

IV. Training Data and Procedure

We train our network with both synthetic and real data. To generate synthetic data, we use 3500 3DMM configurations from the 3DMM fits provided in the 3DDFA dataset [57]. We render each 3DMM configuration from 10 different viewpoints using the path-tracing method in the Mitsuba renderer [20]. The views are obtained by keeping the 3D face frontal and sampling the azimuth angle in the range $[-45, 45]$ with increments of 5, elevation angle in the range $[-30, 30]$ with increments of 5, and in-plane rotations in the range $[-20, 20]$ with increments of 5. We use a set of 90 environment maps available online and up to 5 randomly sampled point light positions to simulate varying illumination conditions, as seen in Figure 5, right.

For real data, we use the Multi-PIE dataset [15], which provides images of 337 subjects captured under 15 viewpoints and 19 illumination conditions. In our experiments, we used 9 viewpoints and discarded the views where faces are upside down or close to 90^o profile. Since the Multi-PIE dataset was captured in a restricted lab environment, we further apply color augmentation. Specifically, we select 200 images from the CelebA dataset [28] as reference images. We apply a histogram matching algorithm in the face region of Multi-PIE images to match those CelebA examples. Given a pair of images of the same subject from Multi-PIE, we apply the same color transformation on both images to preserve photo-consistency. We detect landmark points on each Multi-PIE image using the method of Bulat et al. [5], which we subsequently utilize for 3DMM and viewpoint estimation. We show examples in Figure 5, left.

During training, we first train 3DMMNet for regressing 3DMM parameters and viewpoints. Upon convergence, we

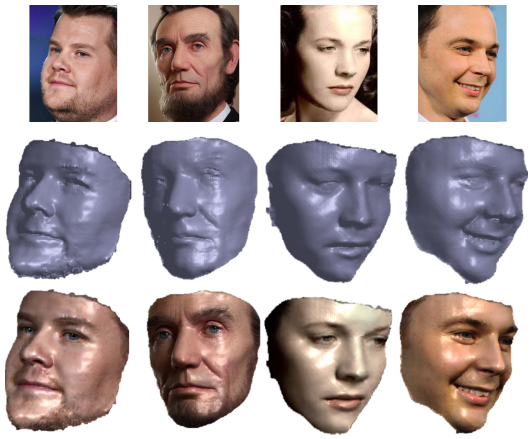

Fig. 6. Qualitative results of our depth prediction network on in the wild faces. Our reconstructions faithfully capture facial geometric features under various pose, expression, etc.

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.

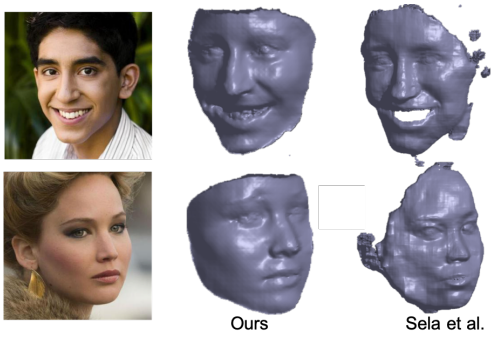

Fig. 7. We compare the depth map predicted by our method and the method of Sela et al. [40]. Our predictions are high quality and better capture the geometric details.

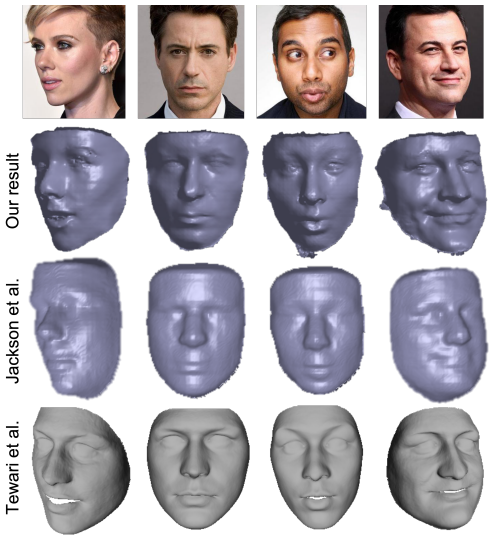

Fig. 8. We compare to state-of-the-art that uses volumetric [19] (row 3) and 3DMM based representation [43] (row 4). While volumetric representations suffer from limited resolution, 3DMMs are not flexible enough to capture geometric details. In contrast, our depth predictions better preserve the identity of the subjects.

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 $50 fps$ on an NVIDIA TitanX GPU, and both the input image and predicted depth map are at resolution 360×360.

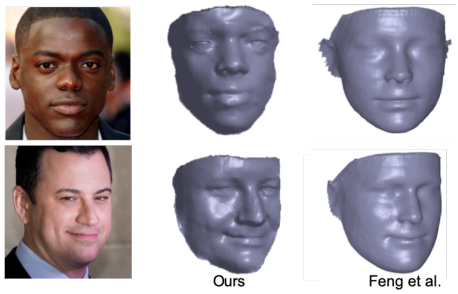

Fig. 9. Comparing to Feng et al.[9] (right), our results (middle) preserves better details of the input face, including the identity, the ethnicity, and the facial expression.

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,

	Mean Err.	Std Err.	Median Err.	90% Err.
[23]	3.89	4.14	2.94	7.34
[58]	3.85	3.23	2.93	7.91
[33]	3.61	2.99	2.72	6.82
[40]	3.51	**2.69**	2.65	6.59
Ours	**2.95**	2.83	**2.22**	**3.26**

TABLE I

QUANTITATIVE EVALUATION ON BU-3DFE.

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.

	AN	DI	FE	HA	NE	SA	SU
[23]	3.47	4.03	3.94	4.30	3.43	3.52	4.19
[58]	4.00	3.93	3.91	3.70	3.76	3.61	3.96
[33]	3.42	3.46	3.64	3.41	4.22	3.59	4.00
[40]	3.67	3.34	3.36	3.01	3.17	3.37	4.41
Ours	**2.96**	**3.00**	**2.85**	**2.77**	**2.67**	**2.95**	**2.75**

TABLE II

THE MEAN ERROR FOR DIFFERENT EXPRESSIONS IN BU-3DFE. LEFT TO RIGHT: ANGER, DISGUST, FEAR, HAPPY, NEUTRAL, SAD, SURPRISE.

The Florence dataset [2] contains 3D face meshes of 53 subjects. Following [19], we render 20 images for each subject in 20 different poses and compute the Normalized Mean Error (average per-vertex Euclidean distance normalized by an outer interocular distance defined in [19]) between our depth prediction and the ground truth 3D shape. ICP is used to compute the closest point correspondence for which we compute the error. As shown in Table III, our results outperform the previous state-of-the-art method using volumetric representation [19] and are comparable to the previous state-of-the-art which used a position map for face reconstruction [9]. However, as we demonstrate in Figure. 9, because of the multi-view weak-supervision used in our system, our reconstruction preserves better facial details compared to the results of [9], which was a fully supervised method trained only on a synthetic dataset.

C. Ablation Studies

Using the same architecture and training losses where applicable, we train two variants of DepthNet with and without data from Multi-PIE. We observe that the model trained with Multi-PIE generalizes better to the real world testing cases (see Figure 10). For example, the shape is more accurate and more robust to pose variation. When trained only with synthetic data, the model is prone to errors such as misalignment between shape and texture. Please see the supplementary material for more ablation experiments.

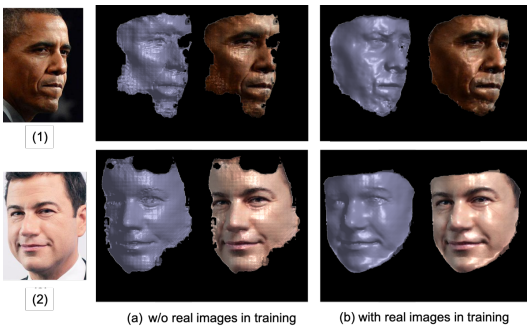

Fig. 10. We compare (a) training DepthNet only using our synthetic dataset (without Multi-PIE) and (b) training with both synthetic and real images. Training with real images enables the model to better generalize to real testing images (notice the misalignment between shape and texture in (a)-first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

	Ours	[19]	[57]	[18]	[9]
NME	0.0475	0.0509	0.0975	0.1253	**0.0362**

TABLE III

NME ON 3D FACE RENDERINGS FROM THE FLORENCE DATASET.

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

[1] B. Amberg, A. Blake, A. Fitzgibbon, S. Romdhani, and T. Vetter. Reconstructing high quality face-surfaces using model based stereo. In *ICCV*, 2007.

[2] A. D. Bagdanov, A. Del Bimbo, and I. Masi. The florence 2d/3d hybrid face dataset. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 79–80. ACM, 2011.

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, pages 187–194, 1999.

[4] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Recognition and Machine Intelligence (PAMI)*, 25(9):1063–1074, 2003.

[5] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017.

[6] C. Cao, M. Chai, O. Woodford, and L. Luo. Stabilized real-time face tracking via a learned dynamic rigidity prior. *ACM Transactions on Graphics (TOG)*, 37(6), nov 2018.

[7] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics (TOG)*, 33(4):43:1–43:10, July 2014.

[8] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 20(3):413–425, 2014.

[9] Y. Feng et al. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018.

[10] R. Garg, B. V. Kumar, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016.

[11] P. Garrido, L. Valgaert, C. Wu, and C. Theobalt. Reconstructing detailed dynamic face geometry from monocular video. *AACM Transactions on Graphics (TOG)*, 32(6), Nov. 2013.

[12] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt. Reconstruction of personalized 3d face rigs from monocular video. *ACM Transactions on Graphics (TOG)*, 35(3):28:1–28:15, May 2016.

[13] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman. Unsupervised training for 3d morphable model regression. In *CVPR*, 2018.

[14] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[15] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image Vision Comput.*, 28(5):807–813, May 2010.

[16] R. A. Güler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, and I. Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. In *CVPR*, 2017.

[17] G. Huang, Z. Liu, L. van der Maaten, and K. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[18] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. J. Christmas, M. Ratsch, and J. Kittler. A multiresolution 3d morphable face model and fitting framework. In *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016.

[19] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *ICCV*, 2017.

[20] W. Jakob. Mitsuba renderer, 2010. http://www.mitsuba-renderer.org.

[21] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *CVPR*, 2016.

[22] A. F. Junyuan Xie, Ross Girshick. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *ECCV*, pages 740–756. Springer, 2016.

[23] I. Kemelmacher-Shlizerman and R. Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Recognition and Machine Intelligence (PAMI)*, 33(2):394–405, 2011.

[24] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollöfer, and C. Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):163, 2018.

[25] H. Kim, M. Zollöfer, A. Tewari, J. Thies, C. Richardt, and C. Theobalt. Inversefacenet: Deep single-shot inverse face rendering from a single image. In *CVPR*, 2018.

[26] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *ArXiv e-prints*, Dec. 2014.

[27] F. Liu, D. Zeng, Q. Zhao, and X. Liu. Joint face alignment and 3d face reconstruction. In *ECCV*, 2016.

[28] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[29] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *Int. Conf. Advanced Video and Signal Based Surveillance*, pages 296–301, 2009.

[30] M. Piotraschke and V. Blanz. Automated 3d face reconstruction from multiple images using quality measures. In *CVPR*, 2016.

[31] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *ECCV*, 2018.

[32] D. J. Rezende, S. M. A. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess. Unsupervised learning of 3d structure from images. In *NIPS*, 2016.

[33] E. Richardson, M. Sela, and R. Kimmel. 3d face reconstruction by learning from synthetic data. In *3DV*, 2016.

[34] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, 2017.

[35] S. Romdhani and T. Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *CVPR*, 2005.

[36] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241, 2015.

[37] J. Roth, Y. Tong, and X. Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *CVPR*, 2016.

[38] M. Sahasrabudhe, Z. Shu, E. Bartrum, R. Alp Guler, D. Samaras, and I. Kokkinos. Lifting autoencoders: Unsupervised learning of a fully-disentangled 3d morphable model using deep non-rigid structure from motion. In *ICCV Workshop*, 2019.

[39] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016.

[40] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *ICCV*, 2017.

[41] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, 2017.

[42] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. Total moving face reconstruction. In *ECCV*, 2014.

[43] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *CVPR*, 2018.

[44] A. Tewari, M. Zollöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and T. Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *ICCV*, 2017.

[45] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *CVPR*, 2016.

[46] A. T. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *CVPR*, 2016.

[47] L. Tran and X. Liu. Nonlinear 3d face morphable model. In *CVPR*, 2018.

[48] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, 2017.

[49] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki. Sfm-net: Learning of structure and motion from video. In *arxiv*, 2017.

[50] S. Wu, H. Huang, T. Portenier, M. Sela, D. Cohen-Or, R. Kimmel, and M. Zwicker. Specular-to-diffuse translation for multi-view reconstruction. In *CVPR*, 2018.

[51] S. Yamaguchi, S. Saito, K. Nagano, Y. Zhao, W. Chen, K. Olszewski, S. Morishima, and H. Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)*, 37(4):162:1–162:14, July 2018.

[52] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *NIPS*. 2016.

[53] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *FGR*, pages 211–216, April 2006.

[54] R. Yu, S. Saito, H. Li, D. Ceylan, and H. Li. Learning dense facial correspondences in unconstrained images. In *ICCV*, 2017.

[55] S. Zafeiriou, A. Roussos, A. Ponniah, D. Dunaway, and J. Booth. Large scale 3d morphable models. *International Journal of Computer Vision*, 2017.

[56] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.

[57] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016.

[58] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *CVPR*, 2015.