Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision

Soubhik Sanyal Timo Bolkart Haiwen Feng Michael J. Black
Perceiving Systems Department
Max Planck Institute for Intelligent Systems
{soubhik.sanyal, timo.bolkart, haiwen.feng, black}@tuebingen.mpg.de

Figure 1: Without 3D supervision, **RingNet** learns a mapping from the pixels of a single image to the 3D facial parameters of the FLAME model [21]. Top: Images are from the CelebA dataset [22]. Bottom: estimated shape, pose and expression.

Abstract

The estimation of 3D face shape from a single image must be robust to variations in lighting, head pose, expression, facial hair, makeup, and occlusions. Robustness requires a large training set of in-the-wild images, which by construction, lack ground truth 3D shape. To train a network without any 2D-to-3D supervision*, we present* RingNet*, which learns to compute 3D face shape from a single image. Our key observation is that an individual's face shape is constant across images, regardless of expression, pose, lighting, etc. RingNet leverages multiple images of a person and automatically detected 2D face features. It uses a novel loss that encourages the face shape to be similar when the identity is the same and different for different people. We achieve invariance to expression by representing the face using the FLAME model. Once trained, our method takes a single image and outputs the parameters of FLAME, which can be readily animated. Additionally we create a new database of faces* "not quite in-the-wild" (NoW) *with 3D head scans and high-resolution images of the subjects in a wide variety of conditions. We evaluate publicly available methods and find that RingNet is* more accurate than methods that use 3D supervision. *The dataset, model, and results are available for research purposes at* http://ringnet.is.tuebingen.mpg.de.

1. Introduction

Our goal is to estimate 3D head and face shape from a single image of a person. In contrast to previous methods, we are interested in more than just a tightly cropped region around the face. Instead, we estimate the full 3D face, head and neck. Such a representation is necessary for applications in VR/AR, virtual glasses try-on, animation, biometrics, etc. Furthermore, we seek a representation that captures the 3D facial expression, factors face shape from expression, and can be reposed and animated. While there have been numerous methods proposed in the computer vision literature to address the problem of facial shape estimation [40], no previous methods address all of our goals.

Specifically, we train a neural network that regresses from image pixels directly to the parameters of a 3D face model. Here we use FLAME [21] because it is more accurate than other models, captures a wide range of shapes, models the whole head and neck, can be easily animated, and is freely available. Training a network to solve this problem, however, is challenging because there is little paired data of 3D heads/faces together with natural images of people. For robustness to imaging conditions, pose, facial hair, camera noise, lighting, etc., we wish to train from a large corpus of in-the-wild images. Such images, by definition, lack controlled ground truth 3D data.

This is a generic problem in computer vision – finding

2D training data is easy but learning to regress 3D from 2D is hard when paired 3D training data is very limited and difficult to acquire. Without ground truth 3D, there are several options but each has problems. Synthetic training data typically does not capture real-world complexity. One can fit a 3D model to 2D image features but this mapping is ambiguous and, consequently, inaccurate. Because of the ambiguity, training a neural network using only a loss between observed 2D, and projected 3D, features does not lead to good results (cf. [17]).

To address the lack of training data, we propose a new method that learns the mapping from pixels to 3D shape *without any supervised 2D-to-3D training data.* To do so, we learn the mapping using only 2D facial features, automatically extracted with OpenPose [29]. To make this possible, our key observation is that multiple images of the same person provide strong constraints on 3D face shape because the shape remains constant although other things may change such as pose, lighting, and expression. FLAME factors pose and shape, allowing our model to learn what is constant (shape) and factor out what changes (pose and expression).

While it is a fact that face shape is constant for an individual across images, we need to define a training approach that lets a neural network exploit this shape constancy. To that end, we introduce *RingNet*. RingNet takes multiple images of a person and enforces that the shape should be similar between all pairs of images, while minimizing the 2D error between observed features and projected 3D features. While this encourages the network to encode the shapes similarly, we find this is not sufficient. We also add to the "ring" a face belonging to a different random person and enforce that the distance in the latent space between all other images in the ring is larger than the distance between the same person. Similar ideas have been used in manifold learning (e.g. triplet loss) [37] and face recognition [26], but, to our knowledge, our approach has not previously been used to learn a mapping from 2D to 3D geometry. We find that going beyond a triplet to a larger ring, is critical in learning accurate geometry.

While we train with multiple images of a person, note that, at run time, we only need a single image. With this formulation, we are able to train a network to regress the parameters of FLAME directly from image pixels. Because we train this with "in the wild" images, the network is robust across a wide range of conditions as illustrated in Fig. 1. The approach is more general, however, and could be applied to other 2D-to-3D learning problems.

Evaluating the accuracy of 3D face estimation methods remains a challenge and, despite many methods that have been published, there are no rigorous comparisons of 3D accuracy across a wide range of imaging conditions, poses, lighting and occlusion. To address this, we collected a new dataset called *NoW (Not quite in-the-Wild)*, with high-resolution ground truth scans and high-quality images of 100 subjects taken in a range of conditions (Fig. 2). NoW is more complex than previous datasets and we use it to evaluate all recent methods with publicly available implementations. Specifically we compare with [34], [35] and [9], which are trained with 3D supervision. Despite not having any 2D-to-3D supervision our RingNet method recovers more accurate 3D face shape. We also evaluate the method qualitatively on challenging in-the-wild face images.

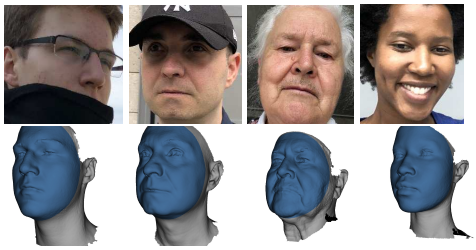

Figure 2: The **NoW dataset** includes a variety of images take in different conditions (top) and high-resolution 3D head scans (bottom). The dark blue region is the part we considered for face challenge.

In summary, the main contributions of our paper are: (1) Full face, head with neck reconstruction from a single face image. (2) RingNet – an end-to-end trainable network that enforces shape consistency across face images of the subject with varying viewing angle, light conditions, resolution and occlusion. (3) A novel shape consistency loss for learning 3D geometry from 2D input. (4) NoW – a benchmark dataset for qualitative and quantitative evaluation of 3D face reconstruction methods. (5) Finally, we make the model, training code, and new dataset freely available for research purposes to encourage quantitative comparison [25].

2. Related work

There are several approaches to the problem of 3D face estimation from images. One approach estimates depth maps, normals, etc.; that is, these methods produce a representation of object shape tied to pixels but specialized for faces. The other approach estimates a 3D shape model that can be animated. We focus on methods in the latter category. In a recent review paper, Zollhöfer et al. [40] describe the state of the art in monocular face reconstruction and provide a forward-looking set of challenges for the field. Note, that the boundary between supervised, weakly supervised, and unsupervised methods is a blurry one. Most methods use some form of 3D shape model, which is learned from scans in advance; we do not call this supervision here. Here the term supervised implies that paired 2D-to-3D data is used; this might be from real data or synthetic data.If a 3D model is first optimized to fit 2D image features, then we

say this uses 2D-to-3D supervision. If 2D image features are used but there is no 3D data in training the network, then this is weakly supervised in general and unsupervised relative to the 2D-to-3D task.

Quantitative evaluation: Quantitative comparison between methods has been limited by a lack of common datasets with complex images and high-quality ground truth. Recently, Feng et al. [10] organized a single image to 3D face reconstruction challenge where they provided the ground truth scans for subjects. Our NoW benchmark is complementary to this method as its focus is on extreme viewing angles, facial expressions, and partial occlusions.

Optimization: Most existing methods require tightly cropped input images and/or reconstruct only a tightly cropped region of the face for which existing shape priors are appropriate. Most current shape models are descendants of the original Blanz and Vetter 3D morphable model (3DMM) [3]. While there are many variations and improvements to this model such as [13], we use FLAME [21] here because both the shape space and expression space are trained from more scans than other methods. Only FLAME includes the neck region in the shape space and models the pose-dependent deformations of the neck with head rotation. Tightly cropped face regions make the estimation of head rotation ambiguous. Until very recently, this has been the dominant paradigm [2, 30, 11]. For example, Kemelmacher-Shlizerman and Seitz [18] use multi-image shading to reconstruct from collection of images allowing changes in viewpoint and shape. Thies et al. [33] achieve accurate results on monocular video sequences. While these approaches can achieve good results with high-realism, they are computationally expensive.

Learning with 3D supervision: Deep learning methods are quickly replacing the optimization-based approaches [35, 39, 19, 16]. For example, Sela et al. [27] use a synthetic dataset to generate an image-to-depth mapping and a pixel-to-vertex mapping, which are combined to generate the face mesh. Tran et al. [34] directly regress the 3DMM parameters of a face model with a dense network. Their key idea is to use multiple images of the same subject and fit a 3DMM to each image using 2D landmarks. They then take a weighted average of the fitted meshes to use it as the ground truth to train their network. Feng et al. [9] regress from image to a UV position map that records the position information of the 3D face and provides dense correspondence to the semantic meaning of each point on UV space. All the aforementioned methods use some form of 3D supervision like synthetic rendering, optimization-based fitting of a 3DMM, or a 3DMM to generate UV maps or volumetric representation. None of the fitting-based methods produce true ground truth for real world face images, while synthetically generated faces may not generalize well to the real world [31]. Methods that rely on fitting a 3DMM to images using 2D-3D correspondences to create a pseudo ground truth are always limited by the expressiveness of the 3DMM and the accuracy of the fitting process.

Learning with weak 3D supervision: Sengupta et al. [28] learn to mimic a Lambertian rendering process by using a mixture of synthetically rendered images and real images. They work with tightly cropped faces and do not produce a model that can be animated. Genova et al. [12] propose an end-to-end learning approach using a differentiable rendering process. They also train their encoder using synthetic data and its corresponding 3D parameters. Tran and Liu [36] learn a nonlinear 3DMM model by using an analytically differentiable rendering layer and in a weakly supervised fashion with 3D data.

Learning with no 3D supervision: MoFA [32] estimates the parameters of a 3DMM and is trained end-to-end using a photometric loss and an optional 2D feature loss. It is effectively a neural network version of the original Blanz and Vetter model in that it models shape, skin reflectance, and illumination to produce a realistic image that is matched to the input. The advantage of this is that the approach is significantly faster than optimization methods [31]. MoFA estimates a tight crop of the face and produces good looking results but has trouble with extreme expressions. They only perform quantitative evaluation on real images using the FaceWarehouse model as the "ground truth"; this is not an accurate representation of true 3D face shape.

The methods that learn without any 2D-to-3D supervision all explicitly model the image formation process (like Blanz and Vetter) and formulate a photometric loss and typically also incorporate 2D face feature detections with known correspondence to the 3D model. The problem with the photometric loss is that the model of image formation is always approximate (e.g. Lambertian). Ideally, one would like a network to learn not just about face shape but about the complexity of real world images and how they relate to shape. To that end, our RingNet approach uses only the 2D face features and no photometric term. Despite (or because of) this, the method is able to learn a mapping from pixels directly to 3D face shape. This is the least supervised of published methods.

3. Proposed method

The goal of our method is to estimate 3D head and face shape from a single face image I. Given an image, we assume the face is detected, loosely cropped, and approximately centered. During training, our method leverages 2D landmarks and identity labels as input. During inference it uses only image pixels; 2D landmarks and identity labels are not used.

Key idea: The key idea can be summarized as follows: 1) The face shape of a person remains unchanged, even though an image of the face may vary in viewing an-

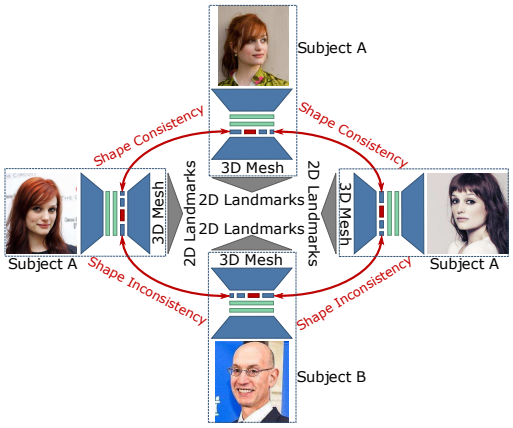

Figure 3: RingNet takes multiple images of the same person (Subject A) and an image of a different person (Subject B) during training and enforces shape consistency between the same subjects and shape inconsistency between the different subjects. The computed 3D landmarks from the predicted 3D mesh projected into 2D domain to compute loss with ground-truth 2D landmarks. During inference, RingNet takes a single image as input and predicts the corresponding 3D mesh. Images are taken from [6]. The figure is a simplified version for illustration purpose.

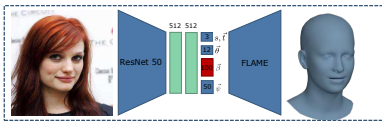

Figure 4: Ring element that outputs a 3D mesh for an image.

gle, lighting condition, resolution, occlusion, expression or other factors. 2) Every person has a unique face shape (not considering identical twins).

We leverage this idea by introducing a shape consistency loss, embodied in our ring-structured network. RingNet (Fig. 3) is a multiple encoder-decoder based architecture, with weight sharing between the encoders, and shape constraints on the shape variables. Each encoder in the ring is a combination of a feature extractor network and a regressor network. Imposing shape constraints on the shape variables forces the network to disentangle facial shape, expression, head pose, and camera parameters. We use FLAME [21] as a decoder to reconstruct 3D faces from the semantically meaningful embedding, and to obtain a decoupling within the embedding space into semantically meaningful parameters (i.e. shape, expression, and pose parameters).

We introduce the FLAME decoder, the RingNet architecture, and the losses in more details in the following.

3.1. FLAME model

FLAME uses linear transformations to describe identity and expression dependent shape variations, and standard linear blend skinning (LBS) to model neck, jaw, and eyeball rotations around $K = 4$ joints. Parametrized by coefficients for shape, $\vec{\beta} \in \mathbb{R}^{|\vec{\beta}|}$, pose $\vec{\theta} \in \mathbb{R}^{3K+3}$, and expression $\vec{\psi} \in \mathbb{R}^{|\vec{\psi}|}$, FLAME returns $N = 5023$ vertices. FLAME models identity dependent shape variations $B_S(\vec{\beta}; \mathcal{S}) : \mathbb{R}^{|\vec{\beta}|} \rightarrow \mathbb{R}^{3N}$, corrective pose blendshapes $B_P(\vec{\theta}; \mathcal{P}) : \mathbb{R}^{3K+3} \rightarrow \mathbb{R}^{3N}$, and expression blendshapes $B_E(\vec{\psi}; \mathcal{E}) : \mathbb{R}^{|\vec{\psi}|} \rightarrow \mathbb{R}^{3N}$ as linear transformations with learned bases $\mathcal{S}$, $\mathcal{E}$, and $\mathcal{P}$. Given a template $\overline{\mathbf{T}} \in \mathbb{R}^{3N}$ in the "zero pose", identity, pose, and expression blendshapes, are modeled as vertex offsets from $\overline{\mathbf{T}}$.

Each of the pose vectors $\vec{\theta} \in \mathbb{R}^{3K+3}$ contains (K+1) rotation vectors in axis-angle representation; i.e. one vector per joint plus the global rotation. The blend skinning function $W(\overline{\mathbf{T}}, \mathbf{J}, \vec{\theta}, \mathcal{W})$ then rotates the vertices around the joints $\mathbf{J} \in \mathbb{R}^{3K}$, linearly smoothed by blendweights $\mathcal{W} \in \mathbb{R}^{K \times N}$. More formally, FLAME is given as

$$M(\vec{\beta}, \vec{\theta}, \vec{\psi}) = W(T_P(\vec{\beta}, \vec{\theta}, \vec{\psi}), \mathbf{J}(\vec{\beta}), \vec{\theta}, \mathcal{W}), \qquad (1)$$

with

$$T_P(\vec{\beta}, \vec{\theta}, \vec{\psi}) = \overline{\mathbf{T}} + B_S(\vec{\beta}; \mathcal{S}) + B_P(\vec{\theta}; \mathcal{P}) + B_E(\vec{\psi}; \mathcal{E}). \quad (2)$$

The joints are defined as a function of $\vec{\beta}$ since different face shapes require different joint locations. We use Equation 1 for decoding our embedding space to generate a 3D mesh of a complete head and face.

3.2. RingNet

The recent advances in face recognition (e.g. [38]) and facial landmark detection (e.g. [4, 29]) have led to large image datasets with identity labels and 2D face landmarks. For training, we assume a corpus of 2D face images I_i, corresponding identity labels c_i, and landmarks k_i.

The shape consistency assumption can be formalized by $\vec{\beta}_i = \vec{\beta}_j, \forall c_i = c_j$ (i.e. the face shape of one subject should remain the same across multiple images) and $\vec{\beta}_i \neq \vec{\beta}_j, \forall c_i \neq c_j$ (i.e. the face shape of different subjects should be distinct). RingNet introduces a ring-shaped architecture that jointly optimizes for shape consistency for an arbitrary number input images in parallel. For details regarding the shape consistency, see Section 3.

RingNet is divided into R ring elements $e_{i=1}^{i=R}$ as shown in Figure 3, where each e_i consists of an encoder and a decoder network (see Figure 4). The encoders share weights across e_i, the decoder weights remain fixed during training. The encoder is a combination of a feature extractor network f_{feat} and regression network f_{reg}. Given an image I_i, f_{feat} outputs a high-dimensional vector, which is

then encoded by f_{reg} into a semantically meaningful vector (i.e., $f_{\text{enc}}(I_i) = f_{\text{reg}}(f_{\text{feat}}(I_i))$). This vector can be expressed as a concatenation of the camera, pose, shape and expression parameters, i.e., $f_{\text{enc}}(I_i) = [\text{cam}_i, \vec{\theta}_i, \vec{\beta}_i, \vec{\psi}_i]$, where $\vec{\theta}_i, \vec{\beta}_i, \vec{\psi}_i$ are FLAME parameters.

For simplicity we omit I in the following and use $f_{\text{enc}}(I_i) = f_{\text{enc},i}$ and $f_{\text{feat}}(I_i) = f_{\text{feat},i}$. The regression network iteratively regresses $f_{\text{enc},i}$ in an iterative error feedback loop [17, 7], instead of directly regressing $f_{\text{enc},i}$ from $f_{\text{feat},i}$. In each iteration step, progressive shifts from the previous estimate are made to reach the current estimate. Formally the regression network takes the concatenated $[f^t_{\text{feat},i}, f^t_{\text{enc},i}]$ as input and gives $\delta f^t_{\text{enc},i}$ as output. Then we update the current estimate by,

$$f_{\text{enc},i}{}^{t+1} = f_{\text{enc},i}{}^{t} + \delta f_{\text{enc},i}{}^{t}. \tag{3}$$

This iterative network performs multiple regression iterations per iteration of the entire RingNet training. The initial estimate is set to $\vec{0}$. The output of the regression network is then fed to the differentiable FLAME decoder network which outputs the 3D head mesh.

The number of ring elements R is a hyper-parameter of our network, which determines the number of images processed in parallel with optimized consistency on the $\vec{\beta}$. RingNet allows to use any combination of images of the same subject and images of different subjects in parallel. However, without loss of generality, we feed face images of the same identity to $\{e_j\}_{j=1}^{j=R-1}$ and different identity to e_R. Hence for each input training batch, each slice consists of $R-1$ images of the same person and one image of another person (see Fig. 3).

3.3. Shape consistency loss

For simplicity let us call two subjects who have same identity label "matched pairs" and two subjects who have different identity labels are "unmatched pairs". A key goal of our work is to make a robust end-to-end trainable network that can produce the same shapes from images of the same subject and different shapes for different subjects. In other words we want to make our shape generators discriminative. We enforce this by requiring matched pairs to have a distance in shape space that is smaller by a margin, η, than the distance for unmatched pairs. Distance is computed in the space of face shape parameters, which corresponds to a Euclidean space of vertices in the neutral pose.

In the RingNet structure, e_j and e_k produce $\vec{\beta}_j$ and $\vec{\beta}_k$, which are matched pairs when $j \neq k$ and $j, k \neq R$. Similarly e_j and e_R produce $\vec{\beta}_j$ and $\vec{\beta}_R$, which are unmatched pairs when $j \neq R$. Our shape constancy term is then

$$\left\|\vec{\beta}_j - \vec{\beta}_k\right\|_2^2 + \eta \leq \left\|\vec{\beta}_j - \vec{\beta}_R\right\|_2^2 \tag{4}$$

Thus we minimize the following loss while training RingNet end-to-end, $L_S =$

$$\sum_{i=1}^{n_b} \sum_{j,k=1}^{R-1} \max(0, \left\|\vec{\beta}_{ij} - \vec{\beta}_{ik}\right\|_2^2 - \left\|\vec{\beta}_{ij} - \vec{\beta}_{iR}\right\|_2^2 + \eta) \tag{5}$$

which is normalized to,

$$L_{\text{SC}} = \frac{1}{n_b \times R} \times L_S \tag{6}$$

where n_b is the batch size for each element in the ring.

3.4. 2D feature loss

Finally we compute the L_1 loss between the ground-truth landmarks provided during the training procedure and the predicted landmarks. Note that we do not directly predict 2D landmarks, but 3D meshes with known topology, from which the landmarks are retrieved.

Given the FLAME template mesh, we define for each OpenPose [29] keypoint the corresponding 3D point in the mesh surface. Note that this is the only place where we provide supervision that connects 2D and 3D. This is done only once. While the mouth, nose, eye, and eyebrow keypoints have a fixed corresponding 3D point (referred to as static 3D landmarks), the position of the contour features changes with head pose (referred to as dynamic 3D landmarks). Similar to [5, 31], we model the contour landmarks as dynamically moving with the global head rotation (see Sup. Mat.). To automatically compute this dynamic contour, we rotate the FLAME template between -20 and 40 degrees to the left and right, render the mesh with texture, run OpenPose to predict 2D landmarks, and project these 2D points to the 3D surface. The resulting trajectories are symmetrically transferred between the left and right side of the face.

During training, RingNet outputs 3D meshes, computes the static and dynamic 3D landmarks for these meshes, and projects these into the image plane using the camera parameters predicted in the encoder output. Henceforth we compute the following L_1 loss between the projected landmarks k_{p_i} and the ground-truth 2D landmarks k_i.

$$L_{\text{proj}} = \|w_i \times (k_{p_i} - k_i)\|_1 \tag{7}$$

where w_i is the confidence score of each ground-truth landmark which is provided by the 2D landmark predictor. We set it to 1 if the confidence is above 0.41 and to 0 otherwise. The total loss L_{tot}, which trains RingNet end-to-end is

$$L_{tot} = \lambda_{SC} L_{\text{SC}} + \lambda_{proj} L_{\text{proj}} + \lambda_{\vec{\beta}} \left\|\vec{\beta}\right\|_2^2 + \lambda_{\vec{\psi}} \left\|\vec{\psi}\right\|_2^2 \tag{8}$$

where the λ are the weights of each loss term and the last two terms regularize the shape and expression coefficients.

Since $B_S(\vec{\beta};\mathcal{S})$ and $B_E(\vec{\psi};\mathcal{E})$ are scaled by the squared variance, the L2 norm of $\vec{\beta}$ and $\vec{\psi}$ represent the Mahalanobis distance in the orthogonal shape and expression space.

3.5. Implementation details

The feature extractor network uses a pre-trained ResNet-50 [15] architecture, also optimized during training. The feature extractor network outputs a 2048 dimensional vector. That serves as input to the regression network. The regression network consists of two fully-connected layers of dimension 512 with ReLu activation and dropout, followed by a final linear fully-connected layer with 159-dimensional output. To this 159-dimensional output vector we concatenate the camera, pose, shape, and expression parameters. The first three elements represent scale and 2D image translation. The following 6 elements are the global rotation and jaw rotation, each in axis-angle representation. The neck and eyeball rotations of FLAME are not regressed since the facial landmarks do not impose any constraints on the neck. The next 100 elements are the shape parameters, followed by 50 expression parameters of FLAME. The differentiable FLAME layer is kept fixed during training. We train RingNet for 10 epochs with a constant learning rate of 1e-4, and use Adam [20] for optimization. The different model parameters are $R = 6$, $\lambda_{SC} = 1$, $\lambda_{proj} = 60$, $\lambda_{\vec{\beta}} = 1e-4$, $\lambda_{\vec{\psi}} = 1e-4$, $\eta = 0.5$. The RingNet architecture is implemented in Tensorflow [1] and will be made publicly available. We use VGG2 Face database [6] as our training dataset which consists of face images and their corresponding labels. We run OpenPose [29] on the database and compute 68 landmark points on the face. OpenPose fails for many cases. After cleaning for the failed cases we have around 800K images with their corresponding labels and facial landmarks for our training corpus. We also consider around 3000 extreme pose images with corresponding landmarks provided by [4]. Since for these extreme images we do not have any labels we replicate each image with random crops and scale for matched pair consideration.

4. Benchmark dataset and evaluation metric

This section introduces our NoW benchmark for the task of 3D face reconstruction from single monocular images. The goal of this benchmark is to introduce a standard evaluation metric to measure the accuracy and robustness of 3D face reconstruction methods under variations in viewing angle, lighting, and common occlusions.

Dataset: The dataset contains 2054 2D images of 100 subjects, captured with an iPhone X, and a separate 3D head scan for each subject. This head scan serves as ground-truth for the evaluation. The subjects are selected to contain variations in age, BMI, and sex (55 female, 45 male).

We categorize the captured data in four challenges; *neutral* (620 images), *expression* (675 images), *occlusion* (528 images) and *selfie* (231 images). *Neutral*, *expression* and *occlusion* contain neutral, expressive, and partially occluded face images of all subjects in multiple views, ranging from frontal view to profile view. *Expression* contains different acted facial expressions such as happiness, sadness, surprise, disgust, and fear. *Occlusion* contain images with varying occlusions from e.g. glasses, sunglasses, facial hair, hats or hoods. For the *selfie* category, participants are asked to take selfies with the iPhone, without imposing constraints on the performed facial expression. The images are captured indoor and outdoor to provide variations of natural and artificial light.

The challenge for all categories is to reconstruct a neutral 3D face given a single monocular image. Note that facial expressions are present in several images, which requires methods to disentangle identity and expression to evaluate the quality of the predicted identity.

Capture setup: For each subject we capture a raw head scan in neutral expression with an active stereo system (3dMD LLC, Atlanta). The multi-camera system consists of six gray-scale stereo camera pairs, six color cameras, five speckle pattern projectors, and six white LED panels. The reconstructed 3D geometry contains about 120K vertices for each subject. Each subject wears a hair cap during scanning to avoid occlusions and scanner noise in the face or neck region due to hair.

Data processing: Most existing 3D face reconstruction methods require a localization of the face. To mitigate the influence of this pre-processing step we provide for each image, a bounding box, that covers the face. To obtain bounding boxes for all images, we first run a face detector on all images [38], and then predict keypoints for each detected face [4]. We manually select 2D landmarks for failure cases. We then expand the bounding box of the landmarks to each side by 5% (bottom), 10% (left and right), and 30% to the top to obtain a box covering the entire face including forehead. For the face challenge, we follow processing protocol similar to [10]. For each scan, the face center is selected, and the scan is cropped by removing everything outside of a specified radius. The selected radius is subject specific computed as $0.7 \times (outer\_eye\_dist + nose\_dist)$ (see Figure 2).

Evaluation metric: Given a single monocular image, the challenge consists of reconstructing a 3D face. Since the predicted meshes occur in different local coordinate systems, the reconstructed 3D mesh is rigidly aligned (rotation, translation, and scaling) to the scan using a set of corresponding landmarks between the prediction and the scan. We further perform a rigid alignment based on the scan-to-mesh distance (which is the absolute distance between each scan vertex and the closest point in the mesh surface) between the ground truth scan, and the reconstructed mesh using the landmarks alignment as initialization. The error

for each image is then computed as the scan-to-mesh distance between the ground truth scan, and the reconstructed mesh. Different errors are then reported including cumulative error plots over all distances, median distance, average distance, and standard deviation.

How to participate: To participate in the challenge, we provide a website [25] to download the test images, and to upload the reconstruction results and selected landmarks for each registration. The error metrics are then automatically computed and returned. Note that we do not provide the ground truth scans to prevent fine-tuning on the test data.

5. Experiments

We evaluate RingNet qualitatively and quantitatively and compare our results with publicly available methods, namely: PRNet (ECCV 2018 [9]), Extreme3D (CVPR 2018 [35]) and 3DMM-CNN (CVPR 2017 [34]).

Quantitative evaluation: We compare methods on [10] and our NoW dataset.

Feng et al. benchmark: Feng et al. [10] describe a benchmark dataset for evaluating 3D face reconstruction from single images. They provide a test dataset, that contains facial images and their 3D ground truth face scans corresponding to a subset of the Stirling/ESRC 3D face database. The test dataset contains 2000 2D neutral face images, including 656 high-quality (HQ) and 1344 low-quality (LQ) images. The high quality images are taken in controlled scenarios and the low quality images are extracted from video frames. The data focuses on neutral faces whereas our data has higher variety in expression, occlusion, and lighting as explained in Section 4.

Recall that the methods we compare with (PRNet, Extreme3D, 3DMM-CNN) use 3D supervision for training whereas our approach does not. PRNet [9] requires a very tightly cropped face region to give good results and performs poorly when given the loosely cropped input image that comes with the benchmark database (see Sup. Mat.). Rather than try to crop the images for PRNet, we run it on the given images and note when it succeeds: it outputs meshes for 918 of the low resolution test images and for 509 of the high-quality images. To be able to compare with PRNet, we run all the other methods only on the 1427 images for which PRNet succeeds.

We compute the error using the method in [10], which computes the distance from ground truth scan points to the estimated mesh surface. Figure 5 (left and middle) show the cumulative error curve for different approaches for the low-quality and high-quality images respectively; RingNet outperforms the other methods. Table 1 reports the mean, standard deviation and median errors.

NoW face challenge: For this challenge we use cropped scans like [10] to evaluate different methods. We first perform a rigid alignment of the predicted meshes to the scans for all the compared methods. Then we compute the scan-to-mesh distance [10] between the predicted meshes and the scans as above. Figure 5 (right) shows the cumulative error curves for the different methods; again RingNet outperforms the others. We provide the mean, median and standard division error in Table 2.

Method	Median (mm)		Mean (mm)		Std (mm)	
	LQ	HQ	LQ	HQ	LQ	HQ
PRNet [9]	1.79	1.60	2.38	2.06	2.19	1.79
Extreme3D [35]	2.40	2.37	3.49	3.58	6.15	6.75
3DMM-CNN [34]	1.88	1.85	2.32	2.29	1.89	1.88
Ours	**1.63**	**1.58**	**2.08**	**2.02**	**1.79**	**1.69**

Table 1: Statistics on Feng et al. [10] benchmark

Method	Median (mm)	Mean (mm)	Std (mm)
PRNet [9]	1.51	1.99	1.90
3DMM-CNN [34]	1.83	2.33	2.05
FLAME-neutral [21]	1.24	1.57	1.34
Ours	**1.23**	**1.55**	**1.32**

Table 2: Statistics for the NoW dataset face challenge.

R	Median (mm)	Mean (mm)	Std (mm)
3	1.25	1.68	1.51
4	1.24	1.67	1.50
5	1.20	1.63	1.48
6	**1.19**	**1.63**	**1.48**

Table 3: Effect of varying number of ring elements R. We evaluate on a validation set described in the ablation study.

Qualitative results: Here we show the qualitative results of estimating a 3D face/head mesh from a single face image on CelebA [22] and MultiPIE dataset [14]. Figure 1 shows a few results for RingNet, illustrating its robustness to expression, gender, head pose, hair, occlusions, etc. We show robustness of our approach under different conditions like lighting, poses and occlusion in Figures 6 and 7. Qualitative comparisons are provided in the Sup. Mat.

Ablation study: Here we provide some motivation for the choice of using a ring architecture in RingNet by comparing different values for R in Table 3. We evaluate these on a validation set that contains 2D images and 3D scans of 10 subjects (six subjects from [8], four from [21]) For each subject we choose one neutral scan and two to four scanner images, reconstruct the 3D meshes for the images, and measure the scan-to-mesh reconstruction error after rigid alignments. The error decreases when using a ring structure with more elements over using a single triplet loss only, but it also increases training time. To make a trade of between time and error, we chose $R = 6$ in our experiments.

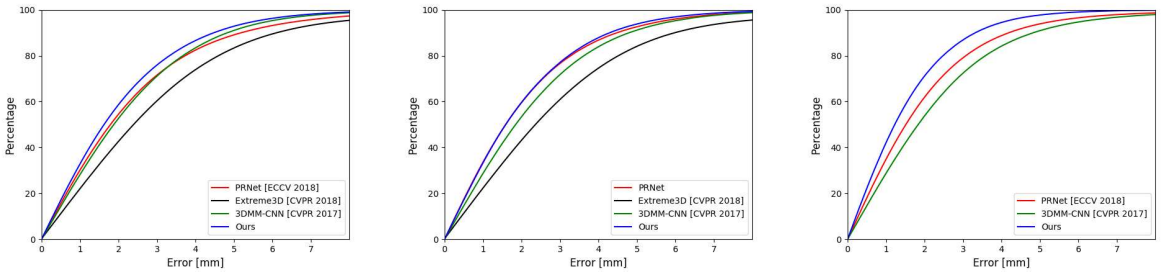

Figure 5: **Cumulative error curves.** Left to right: LQ data of [10]. HQ data of [10]. NoW dataset face challenge.

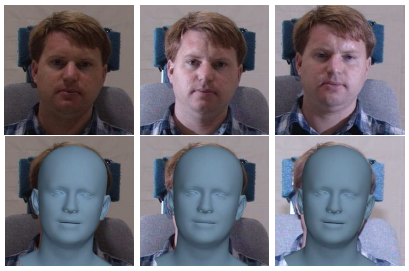

Figure 6: Robustness of RingNet to varying lighting conditions. Images from the MultiPIE dataset [14].

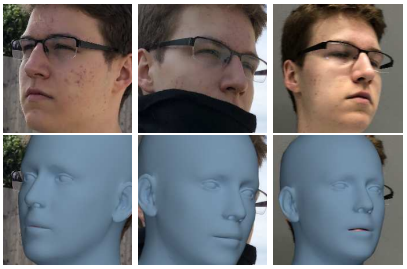

Figure 7: Robustness of RingNet to occlusions, variations in pose, and lighting. Images from the NoW dataset.

6. Conclusion

We have addressed the challenging problem of learning to estimate a 3D, articulated, and deformable shape from a single 2D image with no paired 3D training data. We have applied our RingNet model to faces but the formulation is general. The key idea is to exploit a ring of pairwise losses that encourage the solution to share the same shape for images of the same person and a different shape when they differ. We exploit the FLAME face model to factor face pose and expression from shape so that RingNet can constrain the shape while letting the other parameters vary. Our method requires a dataset in which some of the people appear multiple times, as well as 2D facial features, which can be estimated by existing methods. We provide only the relationship between the standard 2D face features and the vertices of the 3D FLAME model. Unlike previous methods we do not optimize a 3DMM to 2D features, nor do we use synthetic data. Competing methods typically exploit a photometric loss using an approximate generative model of facial albedo, reflectance and shading. RingNet does not need this to learn the relationship between image pixels and 3D shape. In addition, our formulation captures the full head and its pose. Finally, we have created a new public dataset with accurate ground truth 3D head shape and high-quality images taken in a wide range of conditions. Surprisingly, RingNet outperforms methods that use 3D supervision. This opens many directions for future research, for example extending RingNet with [24]. Here we focused on a case with no 3D supervision but we could relax this and use supervision when it is available. We expect that a small amount of supervision would increase accuracy while the large dataset of in-the-wild images provides robustness to illumination, occlusion, etc. Our 2D feature detector does not include the ears, though these are highly distinctive features. Adding 2D ear detections would further improve the 3D head pose and shape. While our model stops with the neck, we plan to extend our model to the full body [23]. It would be interesting to see if RingNet can be extended to reconstruct 3D body pose and shape from images solely using 2D joints. This could go beyond current methods, like HMR [17], to learn about body shape. While RingNet learns a mapping to an existing 3D model of the face, we could relax this and also optimize over the low-dimensional shape space, enabling us to learn a more detailed shape model from examples. For this, incorporating shading cues [32, 28] would help constrain the problem.

Acknowledgement: We thank T. Alexiadis in building the NoW dataset, J. Tesch for rendering results, D. Lleshaj for annotations, A. Osman for supplementary video.

Disclosure: Michael J. Black has received research gift funds from Intel, Nvidia, Adobe, Facebook, and Amazon. He is a part-time employee of Amazon and has financial interests in Amazon and Meshcapade GmbH. His research was performed solely at, and funded solely by, MPI.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016. 6

[2] A. Bas, W. A. Smith, T. Bolkart, and S. Wuhrer. Fitting a 3D morphable model to edges: A comparison between hard and soft correspondences. In *ACCV*, pages 377–391, 2016. 3

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. *In Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 3

[4] 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. 4, 6

[5] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph.*, 33(4):43:1–43:10, July 2014. 5

[6] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018. 4, 6

[7] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *CVPR*, pages 4733–4742, 2016. 5

[8] H. Dai, N. Pears, W. A. P. Smith, and C. Duncan. A 3d morphable model of craniofacial shape and texture variation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 7

[9] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3D face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 2, 3, 7

[10] Z. H. Feng, P. Huber, J. Kittler, P. Hancock, X. J. Wu, Q. Zhao, and M. Rtsch. Evaluation of dense 3D reconstruction from 2D face images in the wild. In *FG*, pages 780–786, 2018. 3, 6, 7, 8

[11] P. Garrido, M. Zollhfer, D. Casas, L. Valgaerts, K. Varanasi, P. Prez, and C. Theobalt. Reconstruction of personalized 3D face rigs from monocular video. *ACM Transactions on Graphics*, 35(3), 2016. 3

[12] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman. Unsupervised training for 3D morphable model regressiond. In *CVPR*, pages 8377–8386, 2018. 3

[13] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schnborn, and T. Vetter. Morphable face models-an open framework. In *FG*, pages 75–82, 2018. 3

[14] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 7, 8

[15] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–645, 2016. 6

[16] 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. 3

[17] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2, 5, 8

[18] I. Kemelmacher-Shlizerman and S. M. Seitz. Face reconstruction in the wild. In *ICCV*, pages 1746–1753, 2011. 3

[19] H. Kim, M. Zollhfer, A. Tewari, J. Thies, C. Richardt, and C. Theobalt. Inversefacenet: Deep monocular inverse face rendering. In *CVPR*, pages 4625–46342, 2018. 3

[20] D. Kinga and J. B. Adam. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, volume 5, 2015. 6

[21] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36(6):194, 2017. 1, 3, 4, 7

[22] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, Dec. 2015. 1, 7

[23] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 8

[24] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black. Generating 3D faces using convolutional mesh autoencoders. In *European Conference on Computer Vision (ECCV)*, volume Lecture Notes in Computer Science, vol 11207, pages 725–741. Springer, Cham, Sept. 2018. 8

[25] RingNet. http://ringnet.is.tuebingen.mpg.de. In *CVPR*, 2019. 2, 7

[26] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 2

[27] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *ICCV*, pages 1585–1594, 2017. 3

[28] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. Jacobs. Sfsnet: Learning shape, reflectance and illuminance of faces in the wild. In *CVPR*, 2018. 3, 8

[29] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 2, 4, 5, 6

[30] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. Total moving face reconstruction. In *ECCV*, pages 796–812, 2014. 3

[31] A. Tewari, M. Zollhfer, P. Garrido, F. Bernard, H. Kim, P. Prez, and C. Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *CVPR*, 2018. 3, 5

[32] A. Tewari, M. Zollhfer, H. Kim, P. Garrido, F. Bernard, P. Prez, and C. Theobalt. MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV*, pages 1274–1283, 2017. 3, 8

[33] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niener. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, pages 2387–2395, 2016. 3

[34] 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*, pages 1493–1502, 2017. 2, 3, 7

[35] A. T. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. Medioni. Extreme 3D face reconstruction: Seeing through occlusions. In *CVPR*, pages 3935–3944, 2018. 2, 3, 7

[36] L. Tran and L. X. Nonlinear 3D face morphable model. In *CVPR*, 2018. 3

[37] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, pages 1473–1480, 2006. 2

[38] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. Sfd: Single shot scale-invariant face detector. 2017. 4, 6

[39] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *CVPR*, pages 146–155, 2016. 3

[40] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Prez, M. Stamminger, M. Niener, and C. Theobalt. State of the art on monocular 3D face reconstruction, tracking, and applications. *Computer Graphics Forum*, 37(2):523–550, 2018. 1, 2