# Semi-Supervised Monocular 3D Face Reconstruction With End-to-End Shape-Preserved Domain Transfer

Jingtan Piao Chen Qian Hongsheng Li
CUHK-SenseTime Joint Laboratory, The Chinese University of Hong Kong Sensetime Research
1155116308@link.cuhk.edu.hk qianchen@sensetime.com hsli@ee.cuhk.edu.hk

## Abstract

*Monocular face reconstruction is a challenging task in computer vision, which aims to recover 3D face geometry from a single RGB face image. Recently, deep learning methods have achieved great improvements on monocular face reconstruction. However, for these methods to reach optimal performance, it is paramount to have large-scale training images with ground-truth 3D face geometry, which is generally difficult for human to annotate. To tackle this problem, we propose a semi-supervised monocular reconstruction method, which jointly optimizes a shape-preserved domain-transfer CycleGAN and a shape estimation network. The framework is semi-supervisely trained with 3D rendered images with ground-truth shapes and in-the-wild face images without any extra annotation. The CycleGAN network transforms all realistic images into rendered style and is end-to-end trained in the overall framework. This is the key difference compared with existing CycleGAN-based learning methods, which just used CycleGAN as a separate training sample generator. Novel landmark consistency loss and edge-aware shape estimation loss are proposed for our two networks to jointly solve the challenging face reconstruction problem. Experiments on public face reconstruction datasets demonstrate the effectiveness of our overall method as well as the individual components.*

## 1. Introduction

3D face reconstruction from monocular images aims at recovering 3D facial geometry from 2D face images. This is an important research topic as human faces play a key role in visual perception and image generation. However, it remains a challenging problem and is far from being solved. Unlike 2D facial landmarks, which can be accurately labeled by humans and robustly estimated by computers thanks to the recent advances of Convolution Neural Networks (CNN), the ground-truth 3D geometry of human faces can only be generated by traditional optimization based methods, such as Gaussian Process [4] and Parameterized Spline [11]. The fitting process of the 3D face geometry contains lots of uncertainties. As a consequence, we cannot easily collect as abundant amount of supervised training data as other computer vision tasks, and the reconstruction accuracy is therefore restricted by the limited amount of training data.

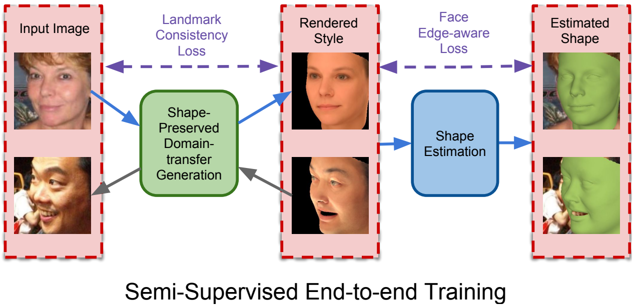

Input Image
Landmark Consistency Loss
Rendered Style
Face Edge-aware Loss
Estimated Shape
Shape-Preserved Domain-transfer Generation
Shape Estimation
Semi-Supervised End-to-end Training

Figure 1. The overall framework of the proposed semi-supervised monocular face reconstruction method.

For face images taken in controlled environments, it is not difficult to use optimization-based methods to obtain their ground-truth 3D face geometry as training data. However, those methods, even with human intervention, cannot guarantee to obtain accurate ground-truth geometry for face images in the wild, where there might exist non-uniform lighting, sided viewing angles, and cluttered backgrounds. Training with such poorly fitted ground truth geometry results in deteriorated reconstruction accuracy.

Several methods have been proposed to tackle the problem of lacking abundant ground-truth face geometry for in-the-wild images. [18] proposed to train a face reconstruction model on only synthetic face images with accurate ground truth, which, however, has difficulty handling realistic images because of the large synthetic-realistic domain gap. Tewari *et al.* [23] proposed a semi-supervised learning approach with photometric loss to supervise the reconstruction for in-the-wild images. But it is likely to overfit to a specific dataset and does not show great generalization capability.

Inspired by the recent research on style transfer and domain adaption with generative adversary network (CycleGAN) [26], we propose a novel semi-supervised deep neural network for 3D face reconstruction from monocular images (see Fig. 1). Our network is trained with both synthetic face images with ground-truth geometry and in-the-wild face images with no extra information. A shape-preserved domain-transfer Cycle-consistent Generative Adversarial Network network is integrated in our framework to bridge the training data from the two image domains. Unlike existing methods [21, 8, 24] that only uses CycleGAN to generate training samples in a separate pre-processing stage, our proposed generative network has three novel distinctions. (1) Instead of mapping synthetic images to realistic ones to create fake training samples, we adopt the generative network to map all realistic images to the rendered domain for face reconstruction. On the one hand, we argue that the images in the rendered domain have accurate ground truth which is not hampered by the translation. On the other, generating realistic images with realistic background, hair, or beard is very difficult. However, removing them from realistic images is relatively easier. (2) Instead of being used as a separate sample generator, our generative network is integrated and end-to-end trained within the face reconstruction network. The realistic-to-synthetic generator is not only required to generate vivid images of rendered style, but also is required to generate images that are easy for the follow-up face shape estimation to recover the face geometry. The co-adpation of our generative network and face shape estimation network is crucial for achieving high reconstruction accuracy.

To effectively train the overall framework for accurate face reconstruction, we also propose novel loss functions. For generative network, we observe that using only the domain discriminators might actually distort the face images, which hinders the final accuracy. We therefore propose a landmark consistency loss to regularize the image translation process maintaining face shapes, which is important for satisfactory domain translation. For training face shape estimation network with in-the-wild images, we propose a novel edge-aware loss to supervise the learning of reconstructing in-the-wild face shapes.

The major contributions of the article can be summarized as three-fold. (1) We propose a novel semi-supervised face reconstruction network with an shape-preserved domain-transfer generative network for better bridging the gap between synthetic and realistic image domains. The generative network is no longer just a training sample generator, but is jointly trained with face reconstruction network to assist accurate face reconstruction. (2) To effectively translate images across the two domains, the shape information of the faces should be maintained as much as possible. We propose a landmark consistency loss as an additional supervision for training the generative model, which provides more constraints to retain the shape information during face translation and thus leads to better reconstruction accuracy. (3) For face reconstruction on images in the wild, other than traditional loss functions on face vertices and face normals, we propose a new loss related to facial-edges on input images. The proposed loss provides better supervisions to train face with larger viewing angles and extreme expressions.

## 2. Releated Works

### 2.1. Monocular 3D Face Reconstruction

By modeling face priors as Gaussian process, Blanz and Vetter [4] proposed 3D Morphable Model (3DMM) to parameterize the shapes of human faces. More accurate models have also been proposed to model more complex facial deformations using blender shapes [7] or skeletons [17]. Those models can model how human faces deforms as smooth surfaces.

Given a set of such parameters, one can obtain a vivid rendered image. However, the inverse process of recovering the 3DMM parameters from monocular images is quite challenging. Other than landmark based optimization methods [3] and differential method based on rendering and rasterization [15], deep learning based methods [27] have been studied recently to recover the face 3DMM parameters. Cao *et al*. [6] proposed a cascaded framework that iteratively refines the recovered face model parameters to handle face images in the wild. However, they have less control on intermedia results. Jackson *et al*. [13] proposed to use volumetric convolution to reconstruct face with greater deformation. However, it introduced heavier computation with volumetric convolution. Feng *et al*. [10] proposed to recover face geometry via predicting a 2D UV position map which records the 3D shape of a complete face in the UV space. It is fast and results in accurate reconstructed shapes with small network structures. However, it has difficulty on handling faces that are taken from extreme viewing angles. Sela *et al*. [20] utilized optimization-based reconstruction methods to make the recovered 3D geometry more stable. It recovers the correspondences between 2D locations in images and UV-coordinates on 3D meshes, and performs ICP registration to obtain the final shape. Although the method shows robust performance on face images of different viewing angles, its network is not end-to-end trainable and requires heavy calculation on optimization.

Other methods explored extra information to assist the learning, including photometric loss [23] , pre-training on synthetic images [18], multi-frame supervisions from videos [22]. These methods are trained to predict a facial texture or color map in additional the 3D face shape to assist face reconstruction. However, they still have difficulty on tackling the sophisticated background and lighting in face images in the wild.

### 2.2. Generative Networks for Cross-domain Training

For image translation, GAN based approaches have shown their great potential. CycleGAN [26] proposed a cross-domain image-to-image translation network, which does not need one-to-one pairs for training. It was used by many approaches [21] for tacking the domain adaption problem by generating domain-transferred images as new training samples with ground-truth labels. However, those methods just used the CycleGAN as a separate training sample generator and there is no supervision or constraint to ensure that the transferred images by the CycleGAN can be correctly recognized by the follow-up classifiers or regressor. In contrast, our proposed generative network is end-to-end trained with the overall framework to guarantee that it serves for the final reconstruction objective. Patch-based discriminators [12] help to concentrate more on distinguishing local textures rather than global image style for better supervising the image-to-image generative networks.

## 3. Proposed Method

The framework of our proposed semi-supervised face reconstruction method consists of a domain-transfer generative network and a 3D shape estimation network. The method utilizes both 3D rendered synthetic images with ground-truth geometry and in-the-wild face images without any annotation for training a robust and accurate face reconstruction network for faces in the wild. The domain-transfer generative network is modeled as a CycleGAN to translate all realistic face images to the rendered style, which is the key difference with existing methods that mostly utilize CycleGANs to generate training samples across different domains. Most importantly, our generative network can be trained with the follow-up 3D shape estimation network in an end-to-end manner to ensure that its main objective is to translate images to optimize face reconstruction. The 3D shape estimation network learns to recover 3D face shapes with ground-truth geometry of the 3D rendered images and realistic face images with a novel face edge-aware loss function.

### 3.1. Face Rendering for Training Data Generation

Since our method is semi-supervised, we first generate 3D rendered images from ground-truth 3D face geometry, which serve as synthetic training data with annotations. In order to generate proper training data, the generation of the rendered images needs to satisfy two requirements. On the one hand, it is important that the synthetic images have ground-truth geometry with enough shape variations to avoid the network being overfitted to some specific face geometry. On the other hand, we should ensure that the styles of the synthetic images are consistent so that the face shape estimation network needs minimal efforts on handling the image styles and can focus on estimating face geometry.

We use a multi-dimensional face generation model, Bessel Face Model [4], which can express precisely most of the faces, to create 3D face shapes with face vertices and face textures in the world coordinate system. The face shapes are controlled by a series of shape, expression, and texture parameters, where the shape parameters control how the pre-defined face bases are linearly combined to generate the face shape. Once we obtain the 3D geometry and texture information of one synthetic face shape, we randomly rotate the faces to simulate different head poses. The 3 rotation parameters, yaw, pitch and roll angles, are randomly chosen from the intervals $[-90°, 90°]$, $[-60°, 60°]$, $[-10°, 10°]$, respectively.

Since real faces may not be expressed as a fully linear combination of the pre-defined face shape bases, in order to synthesize more realistic face deformations, we add slight free-form deformations to some of the generated meshes. A Free-form Deformation is applied to the nose and chin to apply slight changes to face shapes. The bounding grids are manually specified. Then the grid vertices near nose tip and chin tip are moved in the symmetric plane by a distance that follows a Gaussian distribution with standard deviation equaling 0.001 of the grid length.

For face rendering to generate 2D images, we adopt the Phong-Model [5] and project the 3D shape onto a 2D imaging plane. We use parallel lighting from a direction uniformly sampled from the frontal half sphere of the face. The ambient, diffuse, specular components are randomly sampled from Gaussian distributions with white-light mean and standard deviation of $0.01$. The background of the rendered image is set as black, since we expect the domain-transfer generative network (to be introduced) is able to correctly remove the clutter background in the realistic face images.

Following [10], we generate the ground-truth geometry as 2D UV position maps, which record the 3D face shapes in the UV space. To be specific, we use the parameterized UV coordinates as an estimated conformal mapping [9] and then map the mesh boundary to a square. The benefits of adopting the UV map include saving storage compared with volumetric bins, maintaining nearby relationship between neighboring vertices compared with random ordered vertices list, and supporting more flexible deformation compared with expressing in pre-defined facial parameters. An example rendered image, its ground-truth UV map and normal map can seen in Fig. 2.

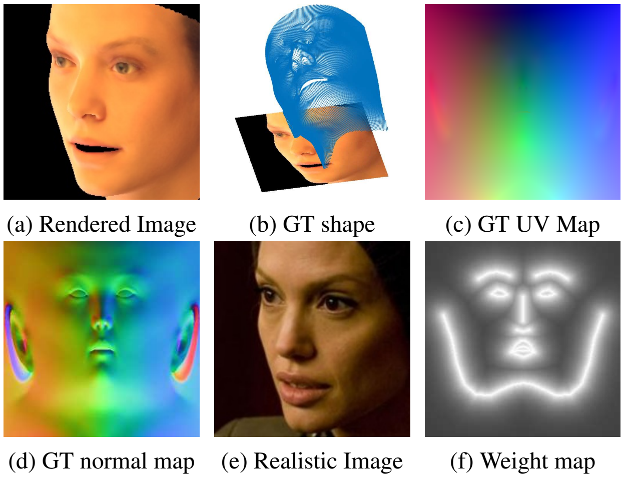

(a) Rendered Image (b) GT shape (c) GT UV Map (d) GT normal map (e) Realistic Image (f) Weight map

Figure 2. (a) A 3D rendered image as described in Section 3.1. (b) The ground-truth geometry of the rendered image. (c) Ground-truth UV map, which encodes 3D face shape in (b) in the UV space and each entry in the map records one vertex's 3D location (shown by a RGB color). (d) Ground-truth normal map of (b). (e) Realistic image corresponding to (a). (f) Weight map for UV estimation with $L2$ loss.

### 3.2. Shape-preserved Domain-transfer Face Image Generation

The domain-transfer generative network in our framework aims to translate all realistic face images in the wild to have the same image style as the 3D rendered face images. The follow-up face shape estimation network will then be process the synthetic images to estimate face 3D geometry. The key difference of the proposed domain-transfer generative network with existing generative methods is that, it is not just a separate cross-domain training sample generator. Instead, our domain-transfer generative network is end-to-end trained in our overall framework to optimize the face reconstruction objective via back-propagation.

Since there are generally no paired ground-truth images for supervising the domain transfer. CycleGAN [26] is adopted as our backbone to convert images back and forth between the realistic image domain and the rendered image domain. Let $R$ and $S$ denote the domains of realistic images and synthetic images, respectively. Two image generators, $F : R \to S$, which translates realistic images to the synthetic domain, and $G : S \to R$, which transfers synthetic images to the realistic domain, are jointly trained for learning domain transferring. In addition, we adopt two patch-based adversarial discriminators $D_S$ and $D_R$ [12], where the former one aims to distinguish whether each patch in the output image is from the synthetic domain or not, and the later one distinguishes realistic-domain samples. For images in the synthetic domain $s \sim p_{data}(s)$, and images in the realistic domain $r \sim p_{data}(r)$, we have the following adversarial objectives,

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(F, D_S, R, S) &= \mathbb{E}_{s\sim p_{data}(s)}[\log D_S(s)] \\ &\quad + \mathbb{E}_{r\sim p_{data}(r)}[\log(1 - D_S(F(r))], \\ \mathcal{L}_{\text{GAN}}(G, D_R, S, R) &= \mathbb{E}_{r\sim p_{data}(r)}[\log D_R(r)] \\ &\quad + \mathbb{E}_{s\sim p_{data}(s)}[\log(1 - D_R(G(s))],\end{aligned} \tag{1}$$

where $F$ and $G$ are optimized to generate images $F(s)$ and $G(r)$ that can fool the two domain discriminators $D_S$ and $D_R$ to perform image-to-image translation across the synthetic and realistic domains. The cycle consistency loss is applied to regularize the two image generators $F$ and $G$ being able to reconstruct the same image after performing the domain transfer twice,

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(F, G) &= \mathbb{E}_{s\sim p_{data}(s)}[\|F(G(s)) - s\|_1] \\ &\quad + \mathbb{E}_{r\sim p_{data}(r)}[\|G(F(r)) - r\|_1].\end{aligned} \tag{2}$$

The above loss functions are similar to those in the classical CycleGAN model, except for the patch-based discriminators rather than the whole image discriminator. However, we observe that with only above mentioned losses, the CycleGAN model cannot guarantee satisfactory domain transferring results. The results usually show undesirable artifacts, and more impotantly, large deformations, which would hinder the following face shape estimation process.

To better regularize the image generation and maintain the faces' 3D shapes, we introduce an additional shape constraint, *i.e.*, after domain transfer, the face geometry should remain the same. However, since we cannot obtain the 3D shapes yet, we relax the constraints to the domain-transferred face images to have the same 2D facial landmarks as their origins. In practice, we utilize a pre-trained and fixed-weight facial landmark estimator network $M$ to estimate the 2D landmarks, which directly regresses the landmark coordinates and can allow errors to back-propagate. We therefore introduce a new landmark consistency loss for cross-domain face image generation,

$$\begin{aligned}\mathcal{L}_{\text{ldmk}}(F, G) &= \mathbb{E}_{r\sim p_{data}(r)}[\|M(r) - M(F(r))\|_2] \\ &\quad + \mathbb{E}_{s\sim p_{data}(s)}[\|M(s) - M(G(s))\|_2],\end{aligned} \tag{3}$$

where $M(\cdot)$ outputs 2D facial landmark coordinates. Note that, although the weights of the landmark estimation network $M$ are fixed, it allows errors to be backpropagation through. Therefore, the errors can be further back-propagated to update the parameters of image generators $F$ and $G$.

The overall objective for our proposed domain-transfer face image generation is

$$\begin{aligned}\mathcal{L}(F, G, D_S, D_R) &= \mathcal{L}_{\text{GAN}}(F, D_S, R, S) \\ &\quad + \mathcal{L}_{\text{GAN}}(G, D_R, S, R) \\ &\quad + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}}(F, G) + \lambda_{\text{ldmk}}\mathcal{L}_{\text{ldmk}}(F, G),\end{aligned} \tag{4}$$

where $\lambda_{\text{cyc}}$ and $\lambda_{\text{ldmk}}$ balance the contributions of cycle consistency loss and our newly proposed landmark consistency loss. With the new landmark consistency loss, we observe significant quality improvements of the domain transferred images. For the generated images in the rendered style, cluttered image background and face accessories can also be automatically removed because such images never appeared in actual synthetic images. See shape-preserved domain-transfer examples in Figs. 4 and 6.

### 3.3. Face Reconstruction with the Edge-aware Loss

Given a face image $r$ in the wild, our domain-transfer generative network $F(r)$ in the above subsection is able to convert it to have the style as 3D rendered images. Then both rendered images $s$ and domain-transferred images $F(r)$ can be processed by our face shape estimation network $E$ without considering their image styles.

**Estimation of shape UV maps.** For a 3D rendered image $s$ that has ground-truth 3D shape associated with it, the network takes the image as input and can be easily trained to minimize the predicted UV map with a weighted $L2$ loss,

$$\mathcal{L}_{\text{uv}}(E) = \mathbb{E}_{s\sim p_{data}(s)}\left[\sum_{i,j} w_{i,j}\|E(s)_{i,j} - E^{gt}(s)_{i,j}\|_2^2\right], \tag{5}$$

where $E(s)$ stands for the network predicted UV map from in the input $s$, $E^{gt}(s)$ denotes the ground-truth face geometry UV map of $s$ generated in Section 3.1, $w$ is a face-region weighting map, and $_{i,j}$ denotes iterating over all entries in the UV and weight maps.

The motivation of adopting the face-region weighting map is based on the observation that not all the vertices on the face model take the same role in controlling the deformation of the surface. Vertices on sharp edges such as face contours, eye lids, nose bridges and mouth lips have more influence. Therefore, a manually designed face-region weighting map indicating the importance of each vertices on the UV-mapping square is generated. Given the important facial landmarks in the UV coordinates, edges are first created to connected the important landmarks. All UV coordinates' deviations from the edges can be modeled as performing the distance transform operation on the edge maps followed by a Gaussian kernel to highlight the face regions near the evident face regions. See one example face weight map in Fig. 3.

**Normal-smoothness Regularization.** We also would like to enhance the smoothness of the estimated face shapes and prevent abrupt curvature changes along the estimated face surface. The estimated shape changes along the $x$- and $y$-dimensions of the UV maps should be perpendicular to the normal maps obtained from the ground-truth face surface,

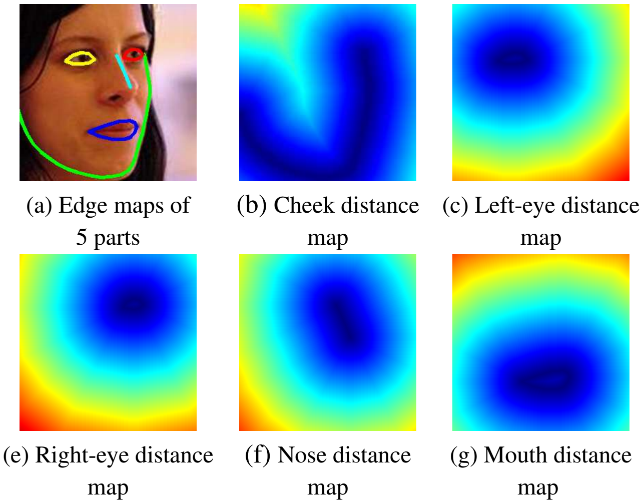
$$\mathcal{L}_{\text{norm}}(E) = \mathbb{E}_{s\sim p_{data}(s)}\Big[ \tag{6}$$
$$\sum_{i,j} |< E(s)_{i+1,j} - E(s)_{i-1,j}, N^{gt}(s)_{i,j} > |+$$
$$\sum_{i,j} |< E(s)_{i,j+1} - E(s)_{i,j-1},\ N^{gt}(s)_{i,j} > |\Big]$$

where $E(s)$ is the estimated shape of the rendered image $s$, $N(s)^{gt}$ is the ground-truth normal map constructed in Section 3.1, and $<,>$ stands for the dot product between the shape-change vectors and the normal vectors at each location (which should be close to zero).

**Unsupervised face shape reconstruction with edge-aware loss.** For a domain-transferred in-the-wild image $F(r)$, however, there is no ground-truth geometry or normal associated with the input image. Existing unsupervised face reconstruction methods mostly utilize the photometric loss to supervise the shape estimation procedure. Those methods first estimate the face shape and re-project it back to the image plane to minimize the differences between original facial pixel values and the re-projected pixel values.

However, calculating such a loss requires simultaneously estimating both shapes and textures and therefore might result in ambiguous shapes. For instance, a planar shape with texture may produce similar results as a slightly bended shapes with warped texture. Therefore, we propose a novel edge-aware loss function that focuses on penalizing the errors of re-projected face edges from the reconstructed face shapes. The key assumption is that, projecting the estimated face shapes to the image plane should result in the same facial landmark locations as those from the original input image. For extracting 2D facial landmarks from the input images, the same fixed-weight facial landmark estimation network $M$ is adopted. The landmarks that delineate cheek, eyes, nose and mouth are connected by edges (see Fig. 3(a)). For each of the 5 facial parts, the distance transform is performed on the edge maps to obtain the distance maps $\{T^1, T^2, \cdots, T^5\}$ (see Fig. 3(b-g)). It has zeros on the landmark edges and the distance values increase as deviating further away from the edges.

On the reconstructed 3D shape, we first locate the same cheek, eye, nose and mouth edge points from the estimated shape UV map. However, sometimes when the face image is captured from a side view, its cheek, nose and mouth edge point locations might change due to self-occlusion. To locate such remaining face edge points corresponding to those from 2D images, we identify 3D edge points as those that are near the original edge point locations and have normal vectors roughly perpendicular to the viewing direction. Such 3D points are denoted as $\{(x_1, y_1, z_1), \cdots, (x_K, y_K, z_K)\}$. By using a sim-


(a) Edge maps of 5 parts

(b) Cheek distance map

(c) Left-eye distance map

(e) Right-eye distance map

(f) Nose distance map

(g) Mouth distance map

Figure 3. 2D face edges and their distance maps for calculating edge-aware loss. (a) 5 important face parts' edges. (b) Cheek edges' distance map. (c) Left-eye edge distance map. (d) Right-eye edge distance map. (f) Nose edge distance map. (g) Mouth edge distance map. Hotter colors represent large distance values.

ple orthogoal projection camera model to remove their $z$-coordinates, they are projected into the image plane as $\Omega = \{(x_1, y_1), \cdots, (x_K, y_K)\}$. We define the following edge-aware loss function,

$$\mathcal{L}_{\text{edge}}(E) = \frac{1}{\sum_{m=1}^{5} |\Omega^m|} \sum_{m=1}^{5} \sum_{(x_k,y_k)\in\Omega^m} T^m_{x_k,y_k}, \quad (7)$$

where $\Omega^m$ is the 2D re-projected edge point set for the $m$th face part, $T_{x_k,y_k}$ denotes the distance value at $(x_k, y_k)$ in the distance map $T$. Intuitively, if the 3D re-projected face edge points well aligns with those landmark edges extracted from the input image, the distance values at the re-projected edge points $T^m_{x_k,y_k}$ should be zeros.

The shape estimation objective can therefore be formulated as

$$\mathcal{L}_{\text{shape}} = \mathcal{L}_{\text{uv}} + \lambda_{\text{norm}}\mathcal{L}_{\text{norm}} + \lambda_{\text{edge}}\mathcal{L}_{\text{edge}}, \quad (8)$$

where $\lambda_{\text{norm}}$ and $\lambda_{\text{edge}}$ weight the contributions of $\mathcal{L}_{\text{norm}}$ and $\mathcal{L}_{\text{edge}}$.

### 3.4. Network Architectures and Training Scheme

**Network structures.** For our image generators $F$ and $G$, we use a U-Net-like network [19], which consists of 8 conv-BN-ReLU blocks as encoder and 8 concat-ReLU-Deconv blocks as decoder. For patch-based discriminators $D_S$ and $D_R$, they have 5 conv-BU-ReLU blocks to downsample the input image to a $8 \times 8$ feature map. A $1 \times 1$ convolution layer is then utilized for patch-based binary classification. For the shape reconstruction network $E$, we design its structure following [10], which is composed of a series of five 2-Residual Block as the encoder and 5 deconv-BN-ReLU as the decoder. A shape UV map is estimated by the estimation network, given the input face image.

**Image pre-processing and augmentation.** For each input image, we first identify its facial landmarks. The face regions are cropped and rotated to have a unified rotation angle and $256 \times 256$ size. During training, after the face normalization, we augment the face image with a random 2D rotation in $[-10°, 10°]$, random $x$- and $y$-translations from a normal distribution $\mathcal{N}(0, 8^2)$. This is to synthesize scenarios where there might be inaccurate landmark locations for face normalization.

**Training scheme.** We train the proposed network in two stages. In the first stage, the domain-transfer generative network with generators $F$ and $G$, and the shape estimation network $E$ are pre-trained independently to obtain good weight initialization. The generative network are pre-trained with unpaired rendered and realistic face images with loss $\mathcal{L}_{\text{gen}}$ and $\lambda_{\text{cyc}} = 0.1$, $\lambda_{\text{ldmk}} = 1$. The shape estimation network is pre-trained with only 3D rendered images and their ground-truth face geometry with loss $\mathcal{L}_{\text{shape}}$ and $\lambda_{\text{norm}} = 0.1$, $\lambda_{\text{edge}} = 0.1$.

In the second stage, the two networks in our framework are end-to-end trained using the ADAM optimizer with both rendered images and in-the-wild realistic images for jointly optimizing the two networks for face reconstruction,

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda_{\text{shape}}\mathcal{L}_{\text{shape}}, \quad (9)$$

where $\lambda_{\text{shape}} = 0.5$ and other weights are kept the same as stage-1. The network is trained with mini-batches of 16 images (8 from each domain) and the learning rate is set to 0.0002.

## 4. Experiments

We test our algorithms on public face reconstruction datasets, including AFLW [14], AFLW-LFPA [27], Florence [1], to evaluate the performance of the proposed semi-supervised face reconstruction method. Not that *no ground-truth face geometry from the the evaluation datasets are used whening training our neural network*. We only utilize 100,000 rendered images based on the Bessel Face Model [4] (as introduced in Section 3.1) with ground-truth geometry and training images without ground-truth from the evaluation datasets for training our model. Therefore, our semi-supervised experimental setup is much more challenging than that of existing methods, which mostly are trained and tested on the same datasets.

### 4.1. Dataset and Evaluation Metrics

**AFLW2000-3D** [14] contains the first 2,000 images from the AFLW dataset, each of which has 68 3D landmarks for face reconstruction evaluation. Since its ground-truth 3DMM parameters are obtained from optimization-based

| Method | Train set | 0-30 | 30-60 | 60-90 | Mean |
|---|---|---|---|---|---|
| 3DDFA [27] | 300W (w/ AFLW) | 3.78 | 4.54 | 7.93 | 5.42 |
| 3DDFA+SDM [27] | 300W (w/ AFLW) | 3.43 | 4.24 | 7.17 | 4.94 |
| Yu *et al.* [25] | 300W+Synthetic | 3.62 | 6.06 | 9.56 | - |
| 3DSTN [2] | 300WLP | 3.15 | 4.33 | 5.98 | 4.49 |
| PRN [10] | 300WLP | **2.75** | **3.51** | 4.61 | **3.62** |
| Ours | Synthetic+Images | 3.56 | 4.06 | **4.11** | 3.88 |

Table 1. 2D NME (%) of 68 landmarks with different yaw angles by compared methods on AFLW2000-3D dataset.

methods, the fitted ground-truth shapes might not be accurate. We therefore measure the Normalized Mean Error (NME) between the 2D re-projected facial landmarks of the estimated 3D face shapes and the ground-truth 3D landmarks for evaluation, *i.e.*,

$$\mathrm{NME}_{2d} = \frac{1}{N}\sum_{n=1}^{N}\frac{\hat{l}_n^{2d} - l_n^{2d}}{\text{box width}}, \tag{10}$$

where $\hat{l}_n^{2d}$ and $l_n^{2d}$ are the re-projected estimated and ground-truth 3D facial landmarks respectively, $N$ is the number of facial landmarks, and "box width" is the width of the face detection box associated with each test image.

**AFLW-LFPA** is another dataset constructed from the AFLW dataset [16], which contains more than 4,000 training images with ground-truth face geometry and 1,299 test images. The Normalized Mean Error (NME) of 3D face vertices can be calculated as

$$\mathrm{NME}_{3d} = \frac{1}{N}\sum_{n=1}^{N}\frac{\hat{v}_n^{3d} - v_n^{3d}}{d}, \tag{11}$$

where $\hat{v}_n^{3d}$ and $v_n^{3d}$ are vertices of the predicted and ground-truth meshes, and $d$ is the 3D interocular distance.

**Florence** [1] is a 3D face dataset that contains 53 subjects with its ground truth 3D mesh acquired from a structured-light scanning system. In our experiments, each subject's face images are generated by renderings with different poses following the same setup as [10]: a pitch of $-15^\circ$, $20^\circ$ and $25^\circ$ degrees and spaced rotations between $-80^\circ$ and $80^\circ$.

## 4.2. Evaluation on AFLW

In Table. 1, we list results of several face reconstruction algorithms' performance on the AFLW-LFPA dataset. We evaluate their average $\mathrm{NME}_{2d}$ on the test set for comparison. Yu *et al.* and 3DSTN [2] are methods that only predict 3D landmarks, which are trained on labeled ground-truth. 3DDFA and 3DDFA+SDM [27] is a method regressing the 3DMM face parameters using a single CNN. PRN [10] also adopts 2D CNN to estimate face UV maps but is trained in a fully-supervised manner.

Our proposed method is able to achieved the 2nd smallest $\mathrm{NME}_{2d}$. This demonstrates our proposed framework's strong capability of using cross-domain data for achieving accurate face reconstruction performance without using any ground-truth geometry in the target domain. For the most extreme poses (yaw angle = $[60^\circ, 90^\circ]$), our algorithm performs best, which is because of the large number of rendered face images with large yaw angles.

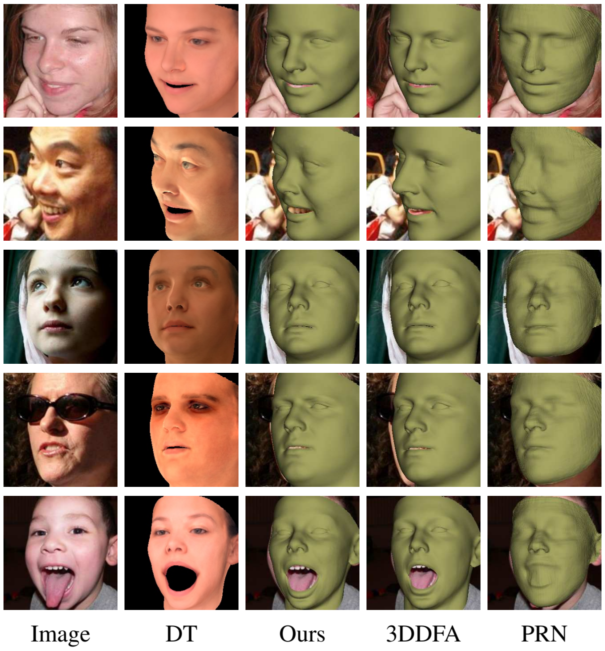


Figure 4. Example face reconstruction results by our proposed method with domain-transfer (DT) generation, 3DDFA [27], and PRN [10] on AFLW-LFPA dataset. Please zoom in to see details.

## 4.3. Evaluation on AFLW-LFPA

On AFLW-LFPA, in contrast to all the other methods trained in a fully-supervised manner, our proposed method are only trained with ground-truth geometry from rendered 3D images and images without any annotation. Average $\mathrm{NME}_{3d}$ curves vs. percentage of test images are shown in Fig. 5(a). Our method achieves the smallest average $\mathrm{NME}_{3d} = 3.703$. In Fig. 4, we show example domain-transfer images and reconstructed face meshes by our proposed method. Our image generator is able to eliminate accessories (such as sunglasses) and generate images of consistent image styles for accurate face reconstruction.

## 4.4. Evaluation on Florence

On Florence dataset, some example reconstructed face meshes are shown in Fig. 6, by different face reconstruction algorithms. Our proposed shape-perserved domain-transfer generator can sucessfully remove the beard from the person's face while maintaining his face shape. Their $\mathrm{NME}_{3d}$ curves w.r.t. percentage of test images are shown in Fig. 5(b). Similar to the AFLW-LFPA datasets, all the compared methods, PRN [10], 3DDFA [27], VRN-

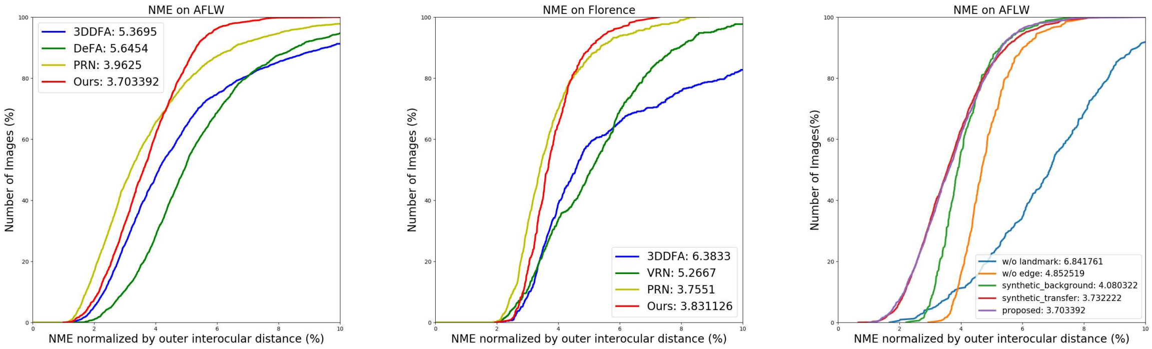


Figure 5. Average $\mathrm{NME}_{3d}$ vs. the percentage of test images on (a) AFLW-LFPA dataset by different compared methods, (b) on Florence dastaset by different compared methods, (c) on AFLW-LFPA dataset by different baseline methods.

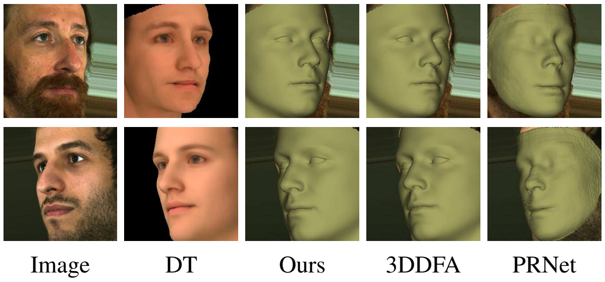


Figure 6. Example reconstruction results by proposed method with domain-transfer (DT) generation, and compared methods 3DDFA [27], PRN [10] on Florence dataset. Please zoom in to see details.

Guided [13] are trained with ground-truth geometry from the Florence dataset. Although we did not use any ground-truth geometry from the Florence dataset. Our method can still achieve comparable average $\mathrm{NME}_{3d}$ with the fully-supervised methods.

## 4.5. Ablation Studies

To evaluate the effectiveness of different components in our framework, we perform ablation studies by changing or removing one certain component from our framework. Their results are reported in Fig. 5(c).

To test the effectiveness of our network. We propose three alternative solutions. 1) Use rendered images with random background instead of black ground (as our method) and directly train the shape estimation network with such synthetic images (denoted as "synthetic+background" in Table 5(c)). Its worse performance denotes that simply training on large-scale synthetic data cannot achieve good performance. The domain gap(especially the artifacts on face edges and hairs) will influence the regression precision. 2) Transfer the image style of the rendered images and utilized the ground-truth meshes of domain-transferred images for training (denoted as "synthetic transfer"). It performs slightly worse than proposed method, denotint that accuracy of the groundtruth and the increase of the training data still cannot handle some important hard cases in reconstruction tasks. 3) Train our generative network and shape estimation network separately and without end-to-end joint optimization. The results show that transfer from sophisticated real images to simplified synthetic images are not easily learned compared with our method with end-to-end training.

We also test removing either $\mathcal{L}_{\text{ldmk}}$ or the edge loss $\mathcal{L}_{\text{edge}}$ . Both methods showed deteriorated accuracies compared with our final model, which demonstrate the effectiveness of the two losses.

## 5. Conclusion

In this paper, we proposed a method that jointly optimizes a shape-preserved domain-transfer generative network and a shape reconstruction network to achieve semi-supervised face reconstruction. Different from existing methods, the proposed domain-transfer generator unifies all input images to have the same style as rendered images. A novel landmark consistency loss is proposed to preserve the original face shapes during translation. The image generative network can be end-to-end trained with the follow-up estimation network to achieve optimal reconstruction accuracy. Such a framework can be trained with only ground-truth geometry from synthetic data and can therefore greatly mitigate the need of large-scale training data for face reconstruction. Extensive experiments demonstrate effectiveness of our proposed face reconstruction method.

## 6. Acknowledgements

This work is supported in part by SenseTime Group Limited, in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants CUHK 14202217, CUHK 14203118, CUHK 14205615, CUHK 14207814, CUHK 14213616, CUHK 14208417, CUHK 14239816, in part by CUHK Direct Grant.

## References

[1] Andrew D Bagdanov, Alberto Del Bimbo, and Iacopo 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.

[2] Chandrasekhar Bhagavatula, Chenchen Zhu, Khoa Luu, and Marios Savvides. Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3980–3989, 2017.

[3] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. Reanimating faces in images and video. In *Computer graphics forum*, volume 22, pages 641–650. Wiley Online Library, 2003.

[4] Volker Blanz and Thomas 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. ACM Press/Addison-Wesley Publishing Co., 1999.

[5] James F Blinn. Models of light reflection for computer synthesized pictures. In *ACM SIGGRAPH computer graphics*, volume 11, pages 192–198. ACM, 1977.

[6] Chen Cao, Qiming Hou, and Kun Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on graphics (TOG)*, 33(4):43, 2014.

[7] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014.

[8] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 994–1003, 2018.

[9] Mathieu Desbrun, Mark Meyer, and Pierre Alliez. Intrinsic parameterizations of surface meshes. In *Computer graphics forum*, volume 21, pages 209–218. Wiley Online Library, 2002.

[10] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. *European Conference on Computer Vision*, 2018.

[11] Hui Guo, Jiayan Jiang, and Liming Zhang. Building a 3d morphable face model by using thin plate splines for face reconstruction. In *Advances in Biometric Person Authentication*, pages 258–267. Springer, 2004.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[13] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1031–1039, 2017.

[14] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4188–4196, 2016.

[15] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2018.

[16] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2144–2151. IEEE, 2011.

[17] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics (TOG)*, 36(6):194, 2017.

[18] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 460–469. IEEE, 2016.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[20] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 1585–1594. IEEE, 2017.

[21] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2107–2116, 2017.

[22] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2549–2559, 2018.

[23] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Pérez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, page 5, 2017.

[24] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018.

[25] Ronald Yu, Shunsuke Saito, Haoxiang Li, Duygu Ceylan, and Hao Li. Learning dense facial correspondences in unconstrained images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4723–4732, 2017.

[26] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.

[27] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.