3D Face Reconstruction from A Single Image Assisted by 2D Face Images in the Wild

Xiaoguang Tu, Jian Zhao *Student Member*, Mei Xie*, Zihang Jiang, Akshaya Balamurugan, Yao Luo, Yang Zhao Lingxiao He, Zheng Ma*, and Jiashi Feng *Member*

Fig. 1: Dense face alignment (odd rows) and 3D face reconstruction (even rows) results from our proposed method. For alignment, only 68 key points are plotted for clear display; for 3D reconstruction, reconstructed shapes are rendered with head light for better view. Our method offers strong robustness and good performance even in presence of large poses (the 4th and 5th columns) and occlusions (the 6th and 7th columns). Best viewed in color.

***Abstract*—3D face reconstruction from a single image is an important task in many multimedia applications. Recent works typically learn a CNN-based 3D face model that regresses coefficients of a 3D Morphable Model (3DMM) from 2D images to perform 3D face reconstruction. However, the shortage of training data with 3D annotations considerably limits performance of these methods. To alleviate this issue, we propose a novel 2D-Assisted Learning (2DAL) method that can effectively use "in the wild" 2D face images with noisy landmark information to substantially improve 3D face model learning. Specifically, taking the sparse 2D facial landmark heatmaps as additional information, 2DAL introduces four novel self-supervision schemes that view the 2D landmark and 3D landmark prediction as a self-mapping process, including the landmark self-prediction consistency for 2D and 3D faces respectively, cycle-consistency over the 2D landmark prediction and self-critic over the predicted 3DMM coefficients based on landmark prediction. Using these four self-supervision schemes, 2DAL significantly relieves the demands for the the conventional paired 2D-to-3D annotations and gives much higher-quality 3D face models without requiring any additional 3D annotations. Experiments on AFLW2000-3D, AFLW-LFPA and Florence benchmarks show that our method outperforms state-of-the-arts for both 3D face reconstruction and dense face alignment by a large margin.**

***Index Terms*—3D face reconstruction, face Alignment, self-supervision.**

I. Introduction

3D face analysis is an important task with a variety of potential applications in computer vision and multimedia, such as face identification [1], [2], [3], [4], [5], [6], emotional state detection [7], [8], human-computer interaction based on facial expression and pose [9], [10], facial replacement [35], [11] and facial animation [12], [13]. For facial animation, which is a major user interface component in modern multimedia systems, it can be described as the problem of finding corresponding feature locations in different faces. Although traditional 2D based methods [14], [15], [16] have made great progress in such tasks, their limitations are evident, *i.e.,* the performance are partly influenced because of unseen regions when the face image appears with large pose variations.

Xiaoguang Tu, Mei Xie, Yao Luo and Zheng Ma are with School of information and communication engineering, UESTC, China. Xiaoguang Tu is also with Department of Electrical and Computer Engineering, NUS, Singapore.

Zihang Jiang and Jiashi Feng are with Department of Electrical and Computer Engineering, National University of Singapore, Singapore.

Jian Zhao is with the Institute of North Electronic Equipment, China.

Akshaya Balamurugan is with Pensees Singapore Institute, Pensees Pte Ltd.

Yang Zhao is with National University of Defense Technology.

Lingxiao He is with JD AI Research.

* The corresponding author. mxie@uestc.edu.cn; zma@uestd.edu.cn

Manuscript received April 19, 2005; revised August 26, 2015.

Unlike 2D based methods, 3D face analysis such as 3D face reconstruction and dense face alignment can help to address the challenges of large pose and occlusion.

Traditional face alignment methods [18], [54] are mainly based on optimization algorithms, *e.g,*, iterative closest point [18], to obtain coefficients for the 3D Morphable Model (3DMM) model and render the corresponding 3D faces from a single face image [69], giving both face reconstruction and dense face alignment results. However, such methods are usually time-consuming due to the high optimization complexity and suffer from local optimal solution and bad initialization. Recent works [44], [69], [50], [66], [65] popularly use CNNs to learn to regress the 3DMM coefficients and significantly improve the reconstruction quality and efficiency.

However, obtaining an accurate 3D face CNN regression model (from input 2D images to 3DMM coefficients) **requires a large number of training faces with 3D annotations**, which are expensive to collect and even not achievable in some cases. Existing methods [69], [40], [35] mainly use 300W-LP [69] to learn a 3D face model. However, this dataset is synthesized in lab-environment by a face profiling algorithm, which generally lack diversity in facial appearances, expressions, occlusions and environment conditions, limiting the generalization performance of resulted 3D face regression models. A model trained on such datasets cannot deal well with various potential cases in the wild that are not covered in the training examples. Moreover, referring to Fig. 5, the ground truth in 300W-LP seems problematic, which can lead to inaccurate 3D face model learning. Although some recent works bypass the 3DMM coefficient regression by using an image-to-volume [40] or image-to-image [35] strategy instead, the ground truths still need to be generated from 3DMM using 300W-LP.

Inspired by the observation that a large number of 2D face datasets [20], [51], [46], [56], [60], [22], [23], [24], [25] are available with obtainable 2D landmark annotations, that could provide valuable information for 3D model learning. We propose a novel learning method that **leverages 2D "in the wild" face images to effectively supervise and facilitate the 3D face model learning**, with which the trained 3D face model can perform 3D face reconstruction and dense face alignment well. However, as these 2D images do not have any 3D annotations, it is not straightforward to exploit them in 3D face model learning. To this end, we propose the 2D-Assisted Learning (2DAL) for better usage of the 2D information from "in the wild" face images with two novelties: First, a self-supervision scheme to model the landmark projection as a self-mapping process and punish the consistency error in 2D and 3D space. Second, a self-critic learning to evaluate the consistency between the predicted 3DMM coefficients and the input "in the wild" face images. By leveraging such additional supervisions, our 2DAL is able to regress more precise and robust coefficients for a 3D Morphable Model, hence improving the performance for 3D face reconstruction and alignment.

For the self-supervision, our 2DAL takes the sparse annotated 2D landmarks as input and fully leverages the consistency within the 2D-to-2D and 3D-to-3D self-mapping procedure as supervision. The model is used to recover 2D landmarks from predicted 3D ones via direct 3D-to-2D projection. Meanwhile, the 3D landmarks predicted from the annotated and recovered 2D landmarks via the model are forced to be the same. Additionally, 2DAL also exploits cycle-consistency over the 2D landmark predictions. That is, taking the recovered 2D landmarks as input, the model generates 2D landmarks by projecting its predicted 3D landmarks and these obtained 2D landmarks are made to approach the annotated ones as much as possible. By leveraging the self-supervision derived from 2D face images without 3D annotations, our method could substantially improve the quality of learned 3D face regression model, even with limited 3D samples and no 3D annotations for the 2D samples. For the self-critic learning, it takes as input both the latent representation and 3DMM coefficients of an face image and learns a critic model to evaluate the intrinsic consistency between the predicted 3DMM coefficients and the corresponding face image, offering another supervision for 3D face model learning.

Our proposed method is principled and effective, and able to fully exploit available data resources. As shown in Figure 1, our method can produce 3D reconstruction and face alignment results with strong robustness to large poses and occlusions. Our contributions are summarized as follows:

- We propose a new scheme that aims to fully utilize the abundant "in the wild" 2D face images to assist 3D face model learning. This is new and different from most common practices that pursue to improve 3D face model by collecting more data with 3D annotations for model training.
- We introduce a new method that is able to train 3D face models with 2D face images by self-supervision. The devised multiple forms of self-supervision are effective and data efficient.
- We develop a new self-critic learning based approach which could effectively improve the 3D face model learning procedure and give a better model, even though the 2D landmark annotations are noisy.
- Comparison on the AFLW2000-3D [69], AFLW-LFPA [42] and Florence [71] datasets shows that our method achieves excellent performance on both tasks of 3D face reconstruction and dense face alignment.

II. Related work

The main problems our method aims to solve are 3D face reconstruction and dense face alignment. This section reviews the works in 3D face reconstruction from a single image and face alignment, which are closely related to our tasks.

3D Face Reconstruction Over the years, various approaches have been proposed to tackle the inherently ill-posed problem of 3D face reconstruction from a single image. Algorithms for single view 3D reconstruction can be broadly classified into two following categories: optimization-based and regression-based. Optimization-based approaches make assumptions about the nature of image formation and express them in the form of energy functions. This is possible because faces represent a set of objects that one can collect some

strong priors about. One popular form of such prior is the 3D Morphable Model (3DMM).

In [27], based on the observation that both the geometric structure and the texture of human faces can be approximated by a linear combination of orthogonal basis vectors obtained by PCA over 100 male and 100 female identities, a 3D morphable model to represent the shape and texture of a 3D face. After that, many efforts have been proposed to improve 3DMM modeling mechanism. Most works obtain the 3DMM coefficients by solving the non-linear optimization problem to establish the correspondences of points between a single face image and the canonical 3D face model, including using facial landmarks [70], [47], [64], [29], [41], [37] and using local features [37], [39], [59]. However, these methods heavily rely on the high accuracy of landmarks or other feature points detector, if the detected results are not reliable the final reconstruction results could be drastically affected.

Recently, the 3DMM parameters are estimated from a single face image using CNN as a regressor, as opposed to non-linear optimization. In [44], [69], [57], [58], cascaded CNN structures are used to regress the 3DMM coefficients, which are time-consuming due to multi-stage. Besides, end-to-end approaches [34], [66], [43] are also proposed to directly estimate the 3DMM coefficients in a holistic manner. More recent works use CNNs to directly obtain the reconstructed 3D face bypassing the 3DMM parameters regression. [40] proposes to map the image pixels to a volumetric representation of the 3D facial geometry through CNN-based regression. While their method is not restricted to the 3DMM space any more, it needs a complex network structure and much time to predict the voxel information. To address the issue of lacking 3D annotations, unsupervised methods have also been proposed. In [72], [73], the researchers use facial texture information to regress 3DMM coefficients without the help of training data. Even good performance could be achieved on frontal face images in ideal environments, large-pose and unbalanced illumination would greatly degrade the performance. Furthermore, such methods are usually model-based by making use of a predefined face templates, which tend to result in a limited geometry constrained in model shape space and post-processing to generate 3D mesh from estimated parameters. Even there are approaches that can recover 3D faces without 3D shape basis, they still rely on 3D facial templates. For examples, [76], [75], [74] reconstruct 3D structure by warping the shape of a reference 3D model. [77] reconstructs the 3D face shapes by learning a 3D Thin Plate Spline(TPS) warping function via a deep CNN. Obviously, the reconstructed face geometry from these methods are still restricted by the reference model, which means the structure differs if the template changes.

Face Alignment The 2D face alignment methods aim at locating a sparse set of fiducial facial landmarks, first with the classic Active Appearance Model (AAM) [31], [62], [67] and Constrained Local Model (CLM) [19], [32], [63]. The regression based method have also been proposed to map the discriminative features around landmarks to the desired landmark positions. By utilizing the feedback characteristic that the output (landmark positions) of the regression has an influence on the input (features at landmarks), the cascaded regression [78] cascades a list of weak regressors to reduce the alignment error progressively and reaches the state of the art [79], [70]. Besides traditional models, CNN based methods [48], [55], [28] have also been employed and achieve state-of-the-art performance on 2D landmark localization. However, most of the methods only regress visible landmarks on faces, which are unable to address large pose or occlusion situations where partial face regions are invisible.

With the development of this field, 3D face alignment have been proposed, aiming to fit a 3DMM [69], [52], [36] or register a 3D facial template [61], [33] to a 2D face image, which makes it possible to deal with the invisible points. The original 3DMM fitting method [26] fits the 3D model by minimizing the pixel-wise difference between image and the rendered face model. It is the first method that can address arbitrary poses but suffers from one-minute-per-image computational cost. Later, some methods estimate 3DMM coefficients and then project the estimated 3D landmarks to a 2D space [43], [29], [41], [45], [45], which significantly improves the efficiency. Recently, the task of dense face alignment attracts growing research attention, with a goal of achieving very dense 3D alignment for large pose face images (including invisible parts). In [50], multi-constraints are used to train a CNN model, which jointly estimates the 3DMM coefficient and provides very dense 3D alignment. Later in [17], [68], the researchers directly learn the correspondence between a 2D face image and a 3D template via a deep CNN, while only visible face regions are considered.

Joint reconstruction and alignment Apart from the methods that focus on 3D face reconstruction and face alignment as two separate tasks, there are methods that perform them jointly. In [49], Liu *et al.* jointly solve the tasks for the first time. They propose to iteratively and alternately update two sets of cascaded regressors, one for updating 2D landmarks and the other for 3D face shape updating, where the 3D face shape and the 2D landmarks are correlated via a 3D-to-2D mapping matrix. In a later work [35], the 3D facial geometry is stored into UV position maps and an image-to-image CNN is trained to directly regress the complete 3D facial structure and the dense facial landmarks along with semantic information from a single image. The method has shown great effectiveness on both the tasks, however it needs plenty of 3D annotated images to generate the UV position maps for pixel-wise image generation. Distinct from these methods, we propose to learn the 3DMM coefficients directly by imposing paired (with 3D ground truth) and unpaired (without 3D ground truth) supervisions on a CNN regression model. The coefficients are then used to fit 3DMM to obtain 3D face reconstruction and dense alignment results together.

Another major difference between our method and the previous ones is that we make full use of "in the wild" face images as additional training subset to assist 3D face model training. Generally, CNN-based methods need huge amounts of 3D annotated data for training. However, the face datasets with 3D annotations are very limited. It appears that only the 300W-LP [69] dataset has been widely used for this kind of training. As we have discussed, the 300W-LP is generated by profiling faces of 300W [60] into larger poses, which is

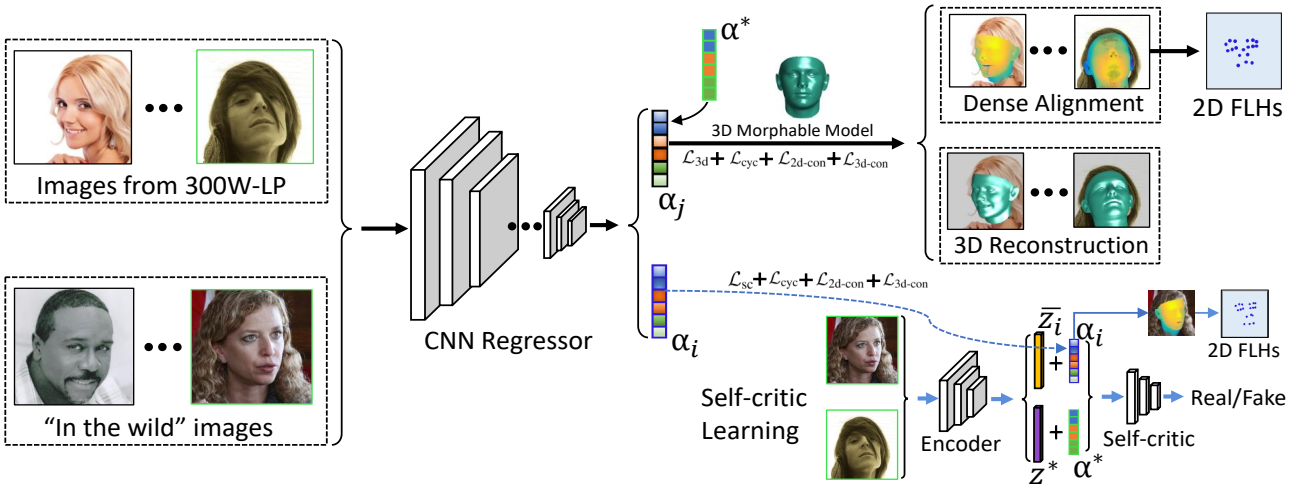


Fig. 2: The pipeline of our 2DAL. It aims to train a CNN regressor model. The model takes as input the face images with 3D annotations (300W-LP) and other images with only 2D annotations ("in the wild"), and predicts coefficients $\boldsymbol{\alpha}_j$ (for 3D annotated images) and $\boldsymbol{\alpha}_i$ (for images only with 2D landmarks) for 3DMM for 3D reconstruction and dense alignment. There are dual training paths. The upper path trains the model through 3D annotation supervision. The bottom path trains the model through self-critic supervision based on the 2D face images. In particular, 2D images are transformed by an encoder to the latent representation, and the self-critic module evaluates whether the predicted coefficients are consistent with the latent representations, by taking ground truth pairs as reference. Best viewed in color.

not strictly unconstrained, and unable to cover all possible scenes in the wild. However, our 2DAL takes advantage of the abundant "in the wild" face images with 2D reliable annotations, to relieve the data deficiency issue in 3D face reconstruction and dense face alignment.

III. Proposed Method

In this section we introduce the proposed 2D-Aided Learning (2DAL) method for simultaneous 3D face reconstruction and dense face alignment. We first revisit the popular 3D morphable model that we adopt to render the 3D faces. Then we explain our method in details, in particular the novel cycle-consistency based self-supervision and the self-critic learning.

A. 3D morphable model

We adopt the 3D Morphable Model (3DMM) [27] to recover the 3D facial geometry from a single face image. The 3DMM renders 3D face shape $\boldsymbol{S} \in \mathbb{R}^{3N}$ that stores 3D coordinates of N mesh vertices with linear combination over a set of PCA basis. We use 40 bases from the Basel Face Model (BFM) [54] to generate the face shape component and 10 bases from the Face Warehouse dataset [30] to generate the facial expression component. The rendering of a 3D face shape is thus formulated as

$$\boldsymbol{S} = \overline{\boldsymbol{S}} + \boldsymbol{A}_{\text{s}}\boldsymbol{\alpha}_{\text{s}} + \boldsymbol{A}_{\text{exp}}\boldsymbol{\alpha}_{\text{exp}},$$

where $\overline{\boldsymbol{S}} \in \mathbb{R}^{3N}$ is the mean shape, $\boldsymbol{A}_{\text{s}} \in \mathbb{R}^{3N\times 40}$ is the shape principle basis trained on the 3D face scans, $\boldsymbol{\alpha}_{\text{s}} \in \mathbb{R}^{40}$ is the shape representation coefficient; $\boldsymbol{A}_{\text{exp}} \in \mathbb{R}^{3N\times 10}$ is the expression principle basis and $\boldsymbol{\alpha}_{\text{exp}} \in \mathbb{R}^{10}$ denotes the corresponding expression coefficient. The target of single-image based 3D face modeling is to predict the coefficients $\boldsymbol{\alpha}_{\text{exp}}$ and $\boldsymbol{\alpha}_{\text{s}}$ for 3D face rendering from a single 2D image.

After obtaining the 3D face shape $\boldsymbol{S}$, it can be projected onto the 2D image plane with the scale orthographic projection to generate a 2D face from specified viewpoint:

$$\boldsymbol{V} = f * \boldsymbol{Pr} * \boldsymbol{\Pi} * \boldsymbol{S} + \boldsymbol{t},$$

where $\boldsymbol{V}$ stores the 2D coordinates of the 3D vertices projected onto the 2D plane, f is the scale factor, $\boldsymbol{Pr}$ is the orthographic projection matrix $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, $\boldsymbol{\Pi}$ is the rotation matrix consisting of 9 parameters, and $\boldsymbol{t}$ is the translation vector. Putting them together, we have in total 62 parameters $\alpha = [f, \boldsymbol{t}, \boldsymbol{\Pi}, \boldsymbol{\alpha}_{\text{s}}, \boldsymbol{\alpha}_{\text{exp}}]$ to regress for the 3D face regressor model.

B. Model overview

As illustrated in Figure 2, the proposed 2DAL model contains 3 modules, *i.e*, a CNN-based *regressor* that predicts 3DMM coefficients from the input 2D image, an *encoder* that transforms the input image into a latent representation, and a *self-critic* that evaluates the input (latent representation, 3DMM coefficients) pairs to be consistent or not.

We use ResNet-50 [38] to implement the CNN regressor. The encoder contains 6 convolutional layers, each followed by a ReLU and a max pooling layer. The critic consists of 4 fully-connected layers with 512, 1024, 1024 and 1 neurons respectively, followed by a softmax layer to output a score on the consistency degree of the input pair. The CNN regressor takes a 3-channel RGB face image as input, this RGB face is annotated with 18 facial landmarks which are encoded into a 1-channel 2D Facial Landmark Heatmap (FLH). We first employ a detector to estimate the coordinates of 18 facial landmarks where the locations corresponding to facial landmarks take the value of 1 and others take the value of -1. The FLH is

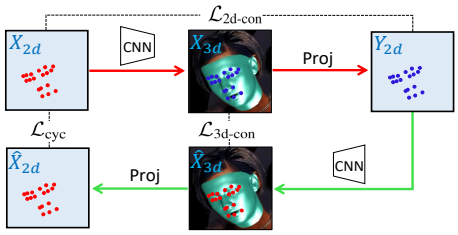


Fig. 3: Illustration on the self-supervision introduced by our 2DAL for utilizing sparse 2D landmark information. The 2D landmark prediction can be viewed as a self-mapping: $X_{2d} \mapsto Y_{2d}$ (forward training) constrained by $\mathcal{L}_{\text{2d-con}}$. To further supervise the model training, we introduce the $\mathcal{L}_{\text{cyc}}$ by mapping back from $Y_{2d} \mapsto \hat{X}_{2d}$ (backward training). The $\mathcal{L}_{\text{3d-con}}$ is employed to constrain landmarks matching in 3D space during the cycle training. Best viewed in color.

filled by a Gaussian distribution where the mean locates at the corresponding coordinate and the standard deviation is set to 2.

Our proposed 2DAL method trains the model using two sets of images, *i.e.,* the images with 3DMM ground truth annotations and the 2D face images with only 2D facial landmark annotations provided by an off-the-shelf facial landmark detector [28]. The model is trained by minimizing the following one conventional 3D-supervision loss and four self-supervision losses.

The 3D-supervision loss is the *weighted coefficient prediction loss* $\mathcal{L}_{\text{3d}}$ over the 3D annotated images that measures how accurate the model can predict 3DMM coefficients. The four self-supervison losses include: 1) the *2D landmark consistency loss* $\mathcal{L}_{\text{2d-con}}$ that measures how well the predicted 3D face shapes can recover the 2D landmark locations for the input 2D images; 2) the *3D landmark consistency loss* $\mathcal{L}_{\text{3d-con}}$; 3) the *cycle consistency loss* $\mathcal{L}_{\text{cyc}}$; 4) the *self-critic loss* $\mathcal{L}_{\text{sc}}$ that estimates the intrinsic consistency between the predicted coefficients and the corresponding face image, conditioned on the face latent representation.

C. Weighted 3DMM coefficient supervision

We deploy the 3DMM ground truth to supervise the model training as in [69], where the contribution of each 3DMM coefficient is re-weighted according to their importance. In this way, the trained model is able to predict closer coefficients $\hat{\alpha}$ to its 3DMM ground truth α^*. Instead of calculating the conventional ℓ_2 loss, we explicitly consider importance of each coefficient and re-weigh their contribution to the loss computation accordingly. Thus we obtain the weighted coefficient prediction loss as follows:

$$\mathcal{L}_{\text{3d}} = (\alpha^* - \hat{\alpha})^\top W (\alpha^* - \hat{\alpha}),$$

where,

$$W = \text{diag}(w_1, \ldots, w_{62}),$$
$$w_i = \frac{1}{\sum_i w_i} \|H(\hat{\alpha}_i) - H(\alpha^*)\|.$$

Here w_i indicates importance of the i^{th} coefficient, computed from how much error it introduces to locations of 2D landmarks after projection. Here $H(\cdot)$ is the sparse landmark projection from rendered 3D shape, α^* is the ground truth and $\hat{\alpha}_i$ is the coefficient whose i^{th} element comes from the predicted coefficient and the others come from α^*. With such a reweighting scheme, during training, the CNN model would first focus on learning the coefficients with larger weight (*e.g.*, the ones for rotation and translation). After decreasing their error and consequently their weights, the model then proceeds to optimize the other coefficients (*e.g.*, the ones for shape and expression).

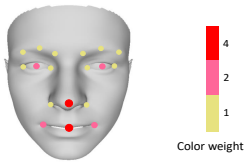


Fig. 4: Illustration of the weight mask used for computing $\mathcal{L}_{\text{2d-con}}$. We assign the highest weight to the red points, the medium weight to the pinky points, the yellow points has the lowest weight. Best viewed in color.

D. 2D assisted learning

To leverage the 2D face images with only annotation of sparse 2D landmark points offered by detector [28], we develop the following self-supervision scheme that offers three different self-supervision losses, including the 2D landmark consistency loss $\mathcal{L}_{\text{2d-con}}$, the 3D landmark consistency loss $\mathcal{L}_{\text{3d-con}}$ and the cycle-consistency loss $\mathcal{L}_{\text{cyc}}$.

Figure 3 gives a systematic overview. The intuition behind this scheme is: if the 3D face estimation model is trained well, it should present consistency in the following three aspects. First, the 2D landmarks Y_{2d} recovered from the predicted 3D landmarks X_{3d} via 3D-2D projection should have small difference with the input 2D landmarks X_{2d}. Second, the predicted 3D landmarks X_{3d} from the input 2D landmarks X_{2d} should be consistent with the 3D landmarks $\hat{X}_{3d}$ recovered from the predicted 2D landmarks Y_{2d} by passing it through the same 3D estimation model. Third, the projected $\hat{X}_{2d}$ from $\hat{X}_{3d}$ should be consistent with the original input X_{2d}, *i.e.*, forming a consistent cycle.

Thus, we define following two landmark consistency losses in our model correspondingly. The $\mathcal{L}_{\text{3d-con}}$ is formulated as

$$\mathcal{L}_{\text{3d-con}} = \sum_{i=1}^{68} \|x_i^{3d} - \hat{x}_i^{3d}\|,$$

where x_i^{3d} is the i^{th} 3D landmark output from the forward pass (see red arrow in Figure 3), $\hat{x}_i^{3d}$ is the i^{th} landmark predicted from the backward pass (see green arrow in Figure 3).

For computing the $\mathcal{L}_{\text{2d-con}}$, we first create a weight mask $V = \{v_1, v_2, ..., v_N\}$ based the contribution of each point. Since the contour landmarks of a 2D face are inaccurate to represent the corresponding points of 3D face, we discard them and sample 18 landmarks from the 68 2D facial landmarks. The weight mask is shown in Figure 4. Here, the mouth center

landmark is the midpoint of two mouth corner points. The $\mathcal{L}_{\text{2d-con}}$ is defined as

$$\mathcal{L}_{\text{2d-con}} = \sum_{i=1}^{18} v_i \times \|x_i^{2d} - y_i^{2d}\|,$$

where x_i^{2d} is the i^{th} 2D landmark of the input face, y_i^{2d} is the i^{th} 2D landmark inferred from the output FLH, and v_i is its corresponding weight. The weight values are specified in Figure 4. We use the following relative weights in our experiments: (red points) : (pinky points) : (yellow points) = 4:2:1 that are set empirically.

We model the 2D facial landmarks prediction as a self-mapping process, and denote $F : X_{2d} \rightarrow Y_{2d}$ as the forward mapping, $Q : Y_{2d} \rightarrow X_{2d}$ as the backward mapping. The backward mapping brings the output landmarks y_i back to its original position x_i, *i.e.*, $x \rightarrow F(x) \rightarrow Q(F(X)) \approx x$. We constrain this mapping using the *cycle consistency loss*:

$$\mathcal{L}_{\text{cyc}} = \mathcal{L}_{\text{2d-con}}(x^{2d}, \hat{x}^{2d}),$$

where x^{2d} are the input 2D facial landmarks, and $\hat{x}^{2d}$ are the landmarks output from $Q(F(X))$.

E. Self-critic learning

We further introduce a self-critic scheme to weakly supervise the model training with the "in the wild" 2D face images. Our self-critic learning adopts similar idea to the conditional GAN (cGAN) [82], [81] that concatenates a condition, which serves as prior knowledge, to generate samples with specific attributes. Given a set of face images $\mathcal{I} = \{I_1, \ldots, I_n\}$ without any 3D annotations and a set of face images $\mathcal{J} = \{(J_1, \alpha_1^*), \ldots, (J_m, \alpha_m^*)\}$ with accurate 3DMM annotations, the CNN regressor model $R : I_i \mapsto \alpha_i$ would output 62 coefficients for each image. Different with cGAN that concatenates the input and output with the same condition to explore the difference between the input and output regarding the given condition, our self-critic module $C(\cdot)$ concatenates each 3DMM coefficients with an exclusive face image representation, to evaluate whether the predicted coefficients are consistent with the input images as the pairs of (J_i, α_i^*). Since each coefficient is closely related to its corresponding face image, the critic model would learn to distinguish the realism of the coefficients conditioned on the latent representation of the input face images. To this end, we feed the input images to an encoder to obtain the latent representation z and then concatenate with their corresponding 3DMM coefficients as the inputs to the critic $C(\cdot)$. It could be difficult for the encoder to encode facial details, but the coarse information such as face shape, rotation angle would not be difficult to learn. Hence evaluating the consistency between the face image and its corresponding 3DMM coefficients can make sense, as the latter also contains such information. The critic is trained in the same way as the adversarial learning by optimizing the following loss:

$$\mathcal{L}_{\text{sc}} = \mathbb{E}_{I \in \mathcal{I}}[\log(C([z^*, \alpha^*])) + \log(1 - C([\overline{z}, R(I)]))],$$

where z^* is the latent representation of a 3D annotated image J, α^* is the 3DMM ground truth, I is the input "in the wild" face image, and $\overline{z}$ is its latent representation.

F. Overall training

The total loss is a weighted sum of the above losses. The parameters of the CNN regression model θ_R, and the self-critic module θ_C are trained alternatively to optimize the following min-max problem:

$$\min_{\theta_R} \max_{\theta_C} \mathcal{L} = \mathcal{L}_{\text{3d}} + \lambda_1 \mathcal{L}_{\text{2d-con}} + \lambda_2 \mathcal{L}_{\text{3d-con}} + \lambda_3 \mathcal{L}_{\text{cyc}} + \lambda_4 \mathcal{L}_{\text{sc}},$$

where λ's are the weighting coefficients for different losses. $\mathcal{L}_{\text{3d}}$ imposes strong paired supervision on the regression model by the annotated 3DMM ground truth. $\mathcal{L}_{\text{2d-con}}$, $\mathcal{L}_{\text{3d-con}}$ and $\mathcal{L}_{\text{cyc}}$ are the consistency losses which leverage the consistency within the 2D-to-2D and 3D-to-3D self-mapping procedure for supervision. The self-critic loss $\mathcal{L}_{\text{sc}}$ encourages the model to output 3D faces that lie on the manifold of human faces, and predict 3DMM coefficients that have the same distribution with the true 3DMM coefficients.

IV. Experiments

In this section, we evaluate our 2DAL qualitatively and quantitatively under various settings for 3D face reconstruction and dense face alignment, respectively.

A. Experimental setup

Our proposed 2DAL is implemented with Pytorch [53]. We use SGD optimizer for the CNN regressor with a learning rate beginning at 5×10^{-5} and decays exponentially. The self-critic uses the Adam as optimizer with the fixed learning rate 1×10^{-4}. The batch size is set as 32. λ_1, λ_2, λ_3 and λ_4 are set as 0.005, 0.005, 1 and 0.005 respectively. We use a two-stage strategy to train our model. In the first stage, we train the model using the overall loss $\mathcal{L}$. In the second stage, we fine-tune our model using the Vertex Distance Cost (VDC), following [69]. The VDC is the Euclidean distance between the reconstructed 3D vertices of the ground-truth coefficients and the reconstructed coefficients. All input images are cropped to the size 120×120.

Training dataset: We train our model on 300W-LP [69], which contains more than 60K face images with annotated 3DMM coefficients. The "in the wild" face images are all from the UMDFaces dataset [20] that contains 367,888 still face images of 8,277 subjects. We use an advanced 2D facial landmark detector [28] to detect the 2D facial landmarks (18 points).

We use the following testing dataset to evaluate our method:

AFLW2000-3D: [69] is constructed by selecting the first 2,000 images from AFLW [46]. Each face is annotated with its corresponding 3DMM coefficients and the 68 3D facial landmarks. We use this dataset to evaluate our method on both 3D face reconstruction and dense face alignment.

AFLW-LFPA: [42] is another extension of AFLW. It is constructed by picking images from AFLW according to the poses. It contains 1,299 test images with a balanced distribution of yaw angle. Each image is annotated with 34 facial landmarks. We use this dataset to evaluate performance for the dense face alignment task. The 34 landmarks are used as the ground truth to measure the accuracy of our results.

Fig. 5: Qualitative results on AFLW2000-3D dataset, which contains the first 2,000 samples of 300W-LP for evaluation. The predictions by 2DAL show that our predictions are more accurate than ground truth in some cases (only 68 points are plotted to show). Green: landmarks predicted by our 2DAL. Red: ground truth from [69]. The thumbnails on the top right corner of each image are the dense alignment results. Best viewed in color.

TABLE I: Performance comparison on AFLW2000-3D and AFLW-LFPA (34 visible landmarks). The NME (%) for faces with different yaw angles are reported. The numbers in bold are the best results on each dataset, the lower is the better. "-" indicates the corresponding result is unavailable.

Methods	AFLW2000-3D (2D Landmakrs)				AFLW2000-3D (3D Landmarks)				AFLW-LFPA
	0° to 30°	30° to 60°	60° to 90°	Mean	0° to 30°	30° to 60°	60° to 90°	Mean	Mean
SDM [52]	3.67	4.94	9.67	6.12	-	-	-	-	-
3DDFA [69]	2.83	3.84	6.33	4.33	7.81	10.48	13.49	10.59	-
3DDFA + SDM [69]	2.78	3.64	6.17	4.20	-	-	-	-	-
PAWF [44]	-	-	-	-	-	-	-	-	4.72
Yu *et al.* [68]	3.62	6.06	9.56	-	-	-	-	-	-
3DSTN [77]	3.15	4.33	5.98	4.49	-	-	-	-	-
DeFA [50]	-	-	-	4.50	10.93	13.10	14.94	12.99	3.86
PRNet [35]	2.75	3.51	4.61	3.62	-	-	-	-	2.93
2DAL (ours)	2.75	3.46	4.45	**3.55**	6.59	8.21	11.22	**8.67**	**2.88**

Florence: [71] 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 generates renderings with different poses as the same with: a pitch of -15, 20 and 25 degrees and spaced rotations between -80 and 80. We compare the performance of our method on face reconstruction against other recent state-of-the-art methods PRN [35], VRN [40] and 3DDFA [69] on this dataset.

B. Evaluation for dense face alignment

We first compare the qualitative results from our method and corresponding ground truths in Figure 5. Although all the state-of-the-art methods of dense face alignment conduct evaluation on AFLW2000-3D, the ground truth of AFLW2000-3D is controversial [28], [68], since its annotation pipeline is based on the Landmarks Marching method in [70]. As can be seen, our results are more accurate than the ground truth in some cases. This is mainly because 2DAL involves a number of the "in the wild" images for training, enabling the model to perform well in cases even unseen in the 3D annotated training data. For fair comparison, we adopt the Normalized Mean Error (NME) [69] as the metric to evaluate the alignment performance. The NME is the mean square error normalized by face bounding box size. The AFLW2000-3D provides a bounding box for face cropping and we use this bounding box size for error normalization. Since some images in AFLW2000-3D contain more than 2 faces, and the face detector sometimes gives wrong face for evaluation (not the test face with ground truth), leading to high NME, we discard the worst 20 cases of each method and only *1,980 images* from AFLW2000-3D are used for evaluation. We evaluate our 2DAL using a sparse set of 68 facial landmarks and also the dense points with both 2D and 3D coordinates, and compare it with other state-of-the-arts, including 3DDFA [69], DeFA [50] and PRNet [35]. The 68 sparse facial landmarks can be viewed as sampling from the dense facial points. Since PRNet is not 3DMM based, and the point cloud of PRN is not corresponding with 3DMM, we only compare with them on the sparse 68 landmarks. The results are shown in Figure 6, we can see our 2DAL achieves the lowest NME (%) on the evaluation of both 2D and 3D coordinates among all the methods. For 3DMM-based methods: 3DDFA and DeFA, our method outperforms them by a large margin on both the 68 spare landmarks and the dense coordinates.

To further investigate performance of our 2DAL across poses and datasets, we report the NME of faces with small, medium and large yaw angles on AFLW2000-3D dataset and the mean NME on both AFLW2000-3D and AFLW-LPFA datasets. The comparison results are shown in Table I. Note that *all the images* from these two datasets are used for evaluation to keep consistent with prior works. The results of the compared method are directly from their published papers. As can be observed, our method achieves the lowest mean NME on both of the two datasets, and the lowest NME across all poses on AFLW2000-3D. Our 2DAL even performs better than PRNet [35], reducing NME (%) by 0.07 and 0.05 on AFLW2000-3D (2D landmarks) and AFLW-

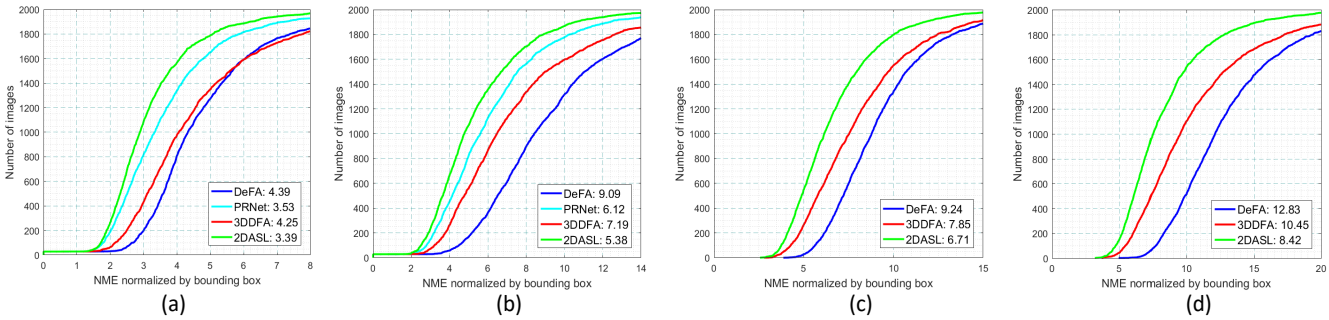


Fig. 6: Error Distribution Curves (EDC) of face alignment results on AFLW2000-3D. The worst 20 cases of each method are discarded. The horizontal axis are the NME (%) in ascending order. The vertical axis are the number of images. Evaluation is performed on the 68 2D landmarks (a), 68 3D landmarks (b), all 2D points (c) and all 3D points (d). The mean NME (%) of each method is shown in the bottom legend.

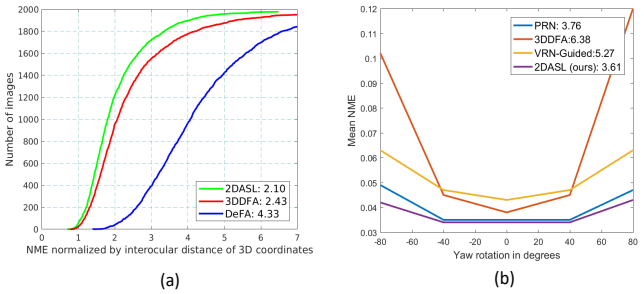


Fig. 7: (a) EDC of face reconstruction results on AFLW2000-3D dataset. (b) Cumulative Error Distribution (CED) curves on Florence dataset with different yaw angles. The NME (%) is normalized by the interocular distance of 3D coordinates.

LFPA, respectively. Especially on large poses (from 60° to 90°), 2DAL achieves 0.16 lower NME (%) than PRNet on the comparison of 2D landmarks on AFLW2000-3D. When 3D dense points (50, 000 + points) is taken for comparison, our 2DAL achieves much better performance than DeFA and 3DDFA, decreasing the mean NME (%) by 4.32 and 1.92 on AFLW2000-3D (3D Landmarks), respectively. We believe more "in the wild" face images used for training would lead to better performance of 2DAL.

C. Evaluation for 3D face reconstruction

As our 2DAL is a 3DMM-based methods, we first evaluate it on the task of 3D face reconstruction on AFLW2000-3D by comparing with other 3DMM-based method, *i.e*, 3DDFA and DeFA. Following [35], we first employ the Iterative Closest Points (ICP) algorithm to find the corresponding nearest points between the reconstructed 3D face and the ground truth point cloud. We then calculate the NME normalized by the interocular distance of 3D coordinates, following the work [35]. Figure 7 (a) shows the comparison results on AFLW2000-3D. As can be seen, the 3D reconstruction results (NME (%)) of 2DAL outperforms 3DDFA by 0.33, and 2.23 for DeFA, which are significant improvements.

To further evaluate the reconstruction performance of our 2DAL across different poses, we calculated the NME for different yaw angle range and compare with 3DDFA [69], VRN-guided [40] and PRNet [35] on Florence dataset. The results of VRN-Guided are copied from [40]. The comparison results are shown in Figure 7 (b), where we can see all the methods perform well in near frontal view, however, 3DDFA and VRN-Guided fail to keep low error as pose becomes large, while PRN and 2DAL keeps relatively stable performance in all pose ranges. However, 2DAL achieves NME (%) 0.15 lower than PRN.

We show some visual results of our 2DAL on AFLW2000-3D and compare them with PRNet and VRN [40] in Figure 8. As VRN-Guided's souce code has not been released, we then compare with another variant of VRN, the VRN-Unguided. It is observed that the reconstructed shape of our 2DAL is more smooth, while both PRNet and VRN-Unguided introduce some artifacts into the reconstructed results, which makes the reconstructed faces look unnatural.

Figure 9 shows the final 3D reconstruction results of 2DAL on Florence dataset, the face texture are sampled from the original 2D face images.

D. Ablation study

We perform ablation study on AFLW2000-3D by evaluating 6 variants of our model: (1) 2DAL (base) which only takes the RGB images as input without self-supervision and self-critic supervision; (2) 2DAL (ss) which takes as input the combination of RGB face images and the corresponding 2D FLHs with self-supervision but without self-critic supervision; (3) 2DAL (sc) which takes as input the RGB face images only using self-critic learning. (4) 2DAL (ss+sc) which contains both the two modules. For each variant, we use the $\mathcal{L}_{\text{2d-con}}$ with (w/) or without (w/o) weight mask.

The ablation study results are shown in Figure 10. As illustrated, adding weights to central points of the facial landmarks reduces the NME (%) of 2DAL (base) by 0.17 and 1.01 on the evaluation of sparse 2D landmarks (68 points) and dense 3D landmarks (50,000+ points), respectively. Similar results have also been observed for other variants, proving that the usage of weight mask to concentrate on reliable key points for learning is effective. In addition to this, both self-critic and self-supervision can improve the performance compared with 2DAL (base). For the variants with weight mask used, the self-critic supervision (2DAL (sc)) decreases NME (%) by 0.03

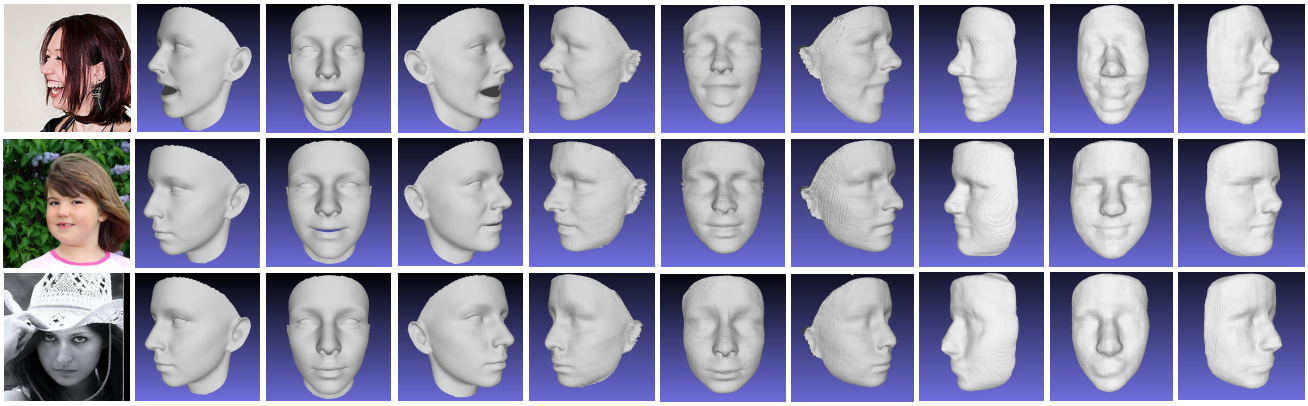


Fig. 8: 3D reconstruction results of 2DAL (columns 2, 3 & 4) , PRNet (columns 5, 6 & 7) and VRN-Unguided (columns 8, 9 & 10). Images of the first column are the original face images.

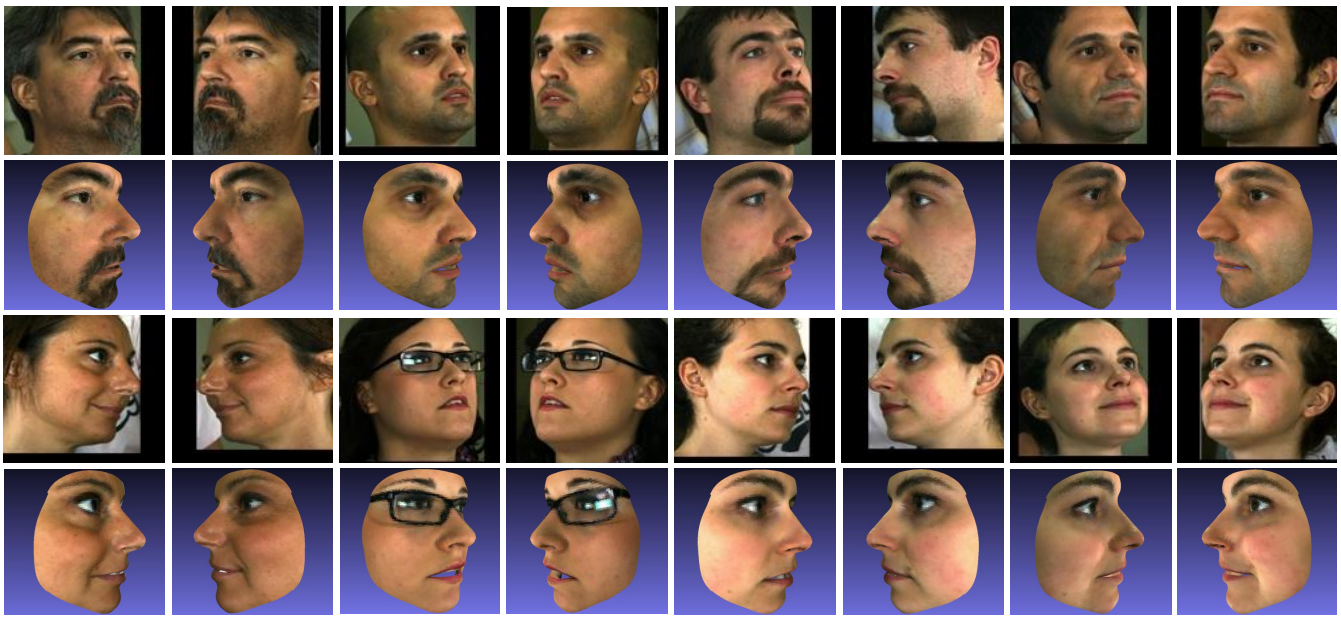


Fig. 9: Qualitative reconstruction results of our 2DAL on Florence dataset, face texture are sampled from the original images. Face regions are cropped for better view.

and 0.17 on 2D landmark evaluation and dense 3D landmark evaluation, respectively. The self-supervision scheme (2DAL (ss)) reduces NME (%) by 0.09 when evaluating on 2D landmarks and 0.48 on dense 3D landmark evaluation. The best results are achieved when self-critic and self-supervision are used together. The NME (%) drops at least 0.23 on 2D facial landmark evaluation, and at least 1.06 on dense 3D landmark evaluation, regardless of the usage of weight mask. It is obvious that the improvement of self-critic is relatively lower than that of self-supervision. Using self-critic and self-supervision together achieve more significant performance than adopting them separately. Because the self-critic adopts the data without paired 3D annotations for training, it can only impose a weak supervision on the model, which cannot guarantee good convergence for the training. However, self-supervision provides strong constraint on the optimization by introducing the self-prediction consistency loss in 2D and 3D reconstruction space with paired 2D annotations. If self-supervision and self-critic are used together during the end-to-end training, they can mutually boost each other to ensure a faster and better convergence for the optimization: the self-supervision ensures the model better optimized to predict more accurate 3DMM coefficients, which are subsequently fed into the self-critic module for better discrimination; the self-critic module evaluates whether the predicted 3D coefficients is consistent with the input images and gives feedback to the regression model to update the parameters for better prediction.

To explore how the performance is affected by the number of "in the wild" face images involved in training, we train our model using different numbers of images. Since the UMDFaces dataset [20] divides the whole dataset into 3 batches, each contains 77,228, 115,126, and 175,534 images respectively. We use the 3 batches and also the whole dataset

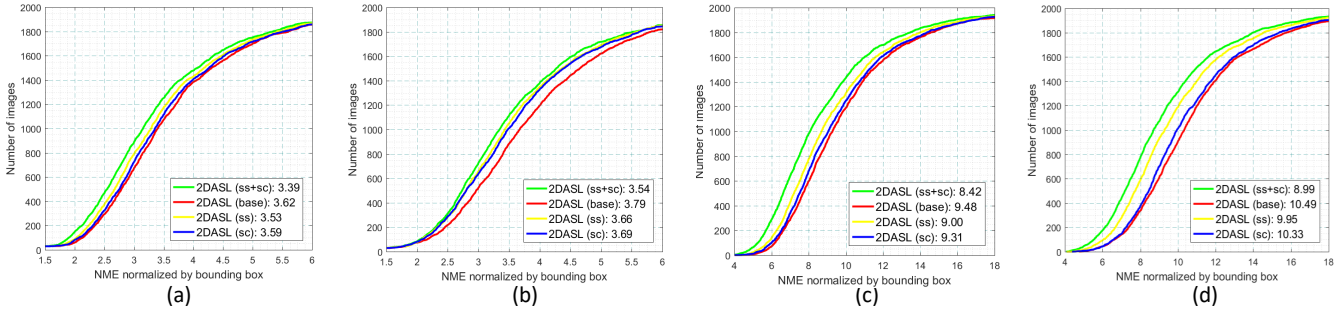


Fig. 10: Abliation study results (NME %). (a) The results of all variants w/ weight mask on 2D facial landmarks; (b) the results of all variants w/o weight mask on 2D facial landmarks; (c) the results of all variants w/ weight mask on dense 3D facial landmarks; (c) the results of all variants w/o weight mask on dense 3D facial landmarks.

TABLE II: The results (NME (%)) of 2DAL, DeFA and 3DDFA by training with different number of "in the wild" face images. "Num. # ITW" indicates the number of the "in the wild" face images used for training. The number in bold is the best result at each training.

Num. # ITW	77,228	115,126	175,534	367,888
3DDFA	4.08	4.01	3.97	3.91
DeFA	4.25	4.13	4.04	3.96
2DAL	**3.83**	**3.64**	**3.52**	**3.39**

TABLE III: The results (NME (%)) of 2DAL by using different weight ratios. The weight ratio of the three subsets are indicate in Figure 4.

Weight setting	1:1:1	2:1:0	4:2:1
Stage 1	4.15	3.75	**3.65**
Stage 2	3.94	3.52	**3.39**

to train our model, respectively. As our model involves more images for training compared with 3DDFA and DeFA that only use 300W-LP for training. For fair comparison, we also use the additional face images from UMDFaces to train 3DDFA and DeFA, in which we minimize the 2D landmark matching error between the input additional images and their reconstructed results. The comparison results are reported in Table II, where we can see the more data that used for aiding training, the lower NME is achieved by the three methods. When using the same number of additional images for training, 2DAL achieves much lower NME (%) than DeFA and 3DDFA, indicating the supervision mechanism of our 2DAL is more advanced than DeFA and 3DDFA.

We also conduct experiments to reveal the influence of our weight mask with different weight ratios, the results are shown in Table III. We could see that weight ratio 1 corresponds to the situation when no weight mask is used, weight ratio 2 and 3 are slightly different on the emphasis in loss function. As can be seen, network without using weight mask has the worst performance compared with other two settings. By adding weights to specific regions such as setting 3 (4:2:1), the performance of the two stages could be significantly improved by 0.5% and 0.55%, respectively.

V. Conclusion

In this paper, we propose a novel 2D-Assisted Learning (2DAL) method for 3D face reconstruction and dense face alignment. The core objective of our method is to make use of "in the wild" 2D face images to substantially improve 3D face model learning, which alleviates the issue of lacking paired 2D-to-3D annotations in 3D face analysis. The sparse 2D facial landmark heatmaps are taken as input of CNN regressor to learn themselves via 3DMM coefficient regression by a deep CNN framework. To supervise and facilitate the 3D face model learning, we introduce two novel self-supervision losses, the self-critic which is employed to weakly supervise the training samples that without 3D annotations, and the landmark consistency loss to constrain landmark matching in both 2D and 3D spaces. Extensive experiments on three challenging datasets show that our 2DAL achieves the best performance for both 3D face reconstruction and dense face alignment by comparing with other state-of-the-art methods.

Acknowledgment

The work of Xiaoguang Tu was partially supported by China Scholarship Council (CSC) grant 201806070011.

References

[1] J. Zhao, L. Xiong, Y. Cheng, Y. Cheng, J. Li, L. Zhou, et al. 3d-aided deep pose-invariant face recognition. In *IJCAI*, Vol. 2, 2018, p. 11.
[2] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, J. Feng. 3d-aided dual-agent gans for unconstrained face recognition. In *IEEE TPAMI* 41 (10) (2018) 2380–2394.
[3] X. Tang, Z. Li, et al. Audio-guided video-based face recognition. In *IEEE TCSVT* 19 (7) (2009) 955–964.
[4] J. Zhao, J. Xing, L. Xiong, S. Yan, J. Feng. Recognizing profile faces by imagining frontal view. In *IJCV* 128 (2) (2020) 460–478.
[5] J. Zhao, L. Xiong, P. K. Jayashree, J. Li, F. Zhao, et al. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *NIPS*, 2017, pp. 66–76.
[6] X. Tang, Z. Li. Video based face recognition using multiple classifiers. In IEEE ICAFGR, 2004, pp. 345–349.
[7] Y. Wang, L. Guan, and A. N. Venetsanopoulos. Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. In *IEEE TMM*, 14(3):597–607, 2012.
[8] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. In *IEEE TPAMI*, 31(1):39–58, 2008.
[9] G. Zen, L. Porzi, E. Sangineto, E. Ricci, and N. Sebe. Learning personalized models for facial expression analysis and gesture recognition. In *IEEE TMM*, 18(4):775–788, 2016.
[10] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan. A deep neural network-driven feature learning method for multi-view facial expression recognition. In *IEEE TMM*, 18(12):2528–2536, 2016.

[11] H. Guo, D. Niu, X. Kong, and X. Zhao. Face replacement based on 2d dense mapping. *ICIGP*, pages 23–28. ACM, 2019.

[12] C. Cao, H. Wu, Y. Weng, T. Shao, and K. Zhou. Real-time facial animation with image-based dynamic avatars. In *ACM TOG*, 35(4), 2016.

[13] Y. Yan, K. Lu, J. Xue, P. Gao, and J. Lyu. Feafa: A well-annotated dataset for facial expression analysis and 3d facial animation. In *arXiv:1904.01509*, 2019.

[14] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, pages 1415–1424, 2017.

[15] S. Liu, X. Yang, Z. Wang, Z. Xiao, and J. Zhang. Real-time facial expression transfer with single video camera. In *CAVW*, 27(3-4):301–310, 2016.

[16] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *ECCV*, pages 818–833, 2018.

[17] R. Alp Guler, G. Trigeorgis, E. Antonakos, P. Snape, et al. Densereg: Fully convolutional dense shape regression in-the-wild. In *CVPR*, pages 6799–6808, 2017.

[18] B. Amberg, S. Romdhani, and T. Vetter. Optimal step nonrigid icp algorithms for surface registration. In *CVPR*, pages 1–8. IEEE, 2007.

[19] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, pages 3444–3451, 2013.

[20] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa. Umdfaces: An annotated face dataset for training deep networks. In *IJCB*, pages 464–473. IEEE, 2017.

[21] C. Bhagavatula, C. Zhu, K. Luu, and M. Savvides. Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In *ICCV*, pages 3980–3989, 2017.

[22] Z. Li, D. Gong, Y. Qiao, D. Tao. Common feature discriminant analysis for matching infrared face images to optical face images. In *IEEE TIP*, 23 (6) (2014) 2436–2445.

[23] Z. Li, D. Gong, Q. Li, D. Tao, X. Li. Mutual component analysis for heterogeneous face recognition. In *ACM TIST*, 7 (3) (2016) 1–23.

[24] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, et al. Towards pose invariant face recognition in the wild. In *CVPR*, 2018, pp. 2207–2216.

[25] J. Zhao, J. Li, X. Tu, F. Zhao, Y. Xin, J. Xing, et al. Multi-prototype networks for unconstrained set-based face recognition. arXiv:1902.04755.

[26] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9):1063–1074, 2003.

[27] V. Blanz, T. Vetter, et al. A morphable model for the synthesis of 3d faces. In *Siggraph*, volume 99, pages 187–194, 1999.

[28] 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*, pages 1021–1030, 2017.

[29] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM TOG*, 33(4):43, 2014.

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

[31] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. Number 6, pages 681–685. IEEE, 2001.

[32] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, volume 1, page 3. Citeseer, 2006.

[33] F. H. de Bittencourt Zavan, A. C. Nascimento, L. P. e Silva, et al. 3d face alignment in the wild: A landmark-free, nose-based approach. In *ECCV*, pages 581–589. Springer, 2016.

[34] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *CVPR*, pages 5908–5917, 2017.

[35] 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*, pages 534–551, 2018.

[36] C. Gou, Y. Wu, F.-Y. Wang, and Q. Ji. Shape augmented regression for 3d face alignment. In *ECCV*, pages 604–615. Springer, 2016.

[37] C. M. Grewe and S. Zachow. Fully automated and highly accurate dense correspondence for facial surfaces. In *ECCV*, pages 552–568. Springer, 2016.

[38] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[39] P. Huber, Z.-H. Feng, W. Christmas, J. Kittler, and M. Rätsch. Fitting 3d morphable face models using local features. In *ICIP*, pages 1195–1199. IEEE, 2015.

[40] 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*, pages 1031–1039, 2017.

[41] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3d face alignment from 2d videos in real-time. In *FG*, volume 1, pages 1–8. IEEE, 2015.

[42] L. A. Jeni, S. Tulyakov, L. Yin, N. Sebe, and J. F. Cohn. The first 3d face alignment in the wild (3dfaw) challenge. In *ECCV*, pages 511–520. Springer, 2016.

[43] A. Jourabloo and X. Liu. Pose-invariant 3d face alignment. In *ICCV*, pages 3694–3702, 2015.

[44] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *CVPR*, pages 4188–4196, 2016.

[45] A. Jourabloo and X. Liu. Pose-invariant face alignment via cnn-based dense 3d model fitting. *IJCV*, 124(2):187–203, 2017.

[46] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCVW*, pages 2144–2151. IEEE, 2011.

[47] Y. J. Lee, S. J. Lee, K. R. Park, J. Jo, and J. Kim. Single view-based 3d face reconstruction robust to self-occlusion. *EURASIP JASP*, 2012(1):176, 2012.

[48] Z. Liang, S. Ding, and L. Lin. Unconstrained facial landmark localization with backbone-branches fully-convolutional networks. *arXiv:1507.03409*, 2015.

[49] F. Liu, D. Zeng, Q. Zhao, and X. Liu. Joint face alignment and 3d face reconstruction. In *ECCV*, pages 545–560. Springer, 2016.

[50] Y. Liu, A. Jourabloo, W. Ren, and X. Liu. Dense face alignment. In *ICCV*, pages 1619–1628, 2017.

[51] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015.

[52] J. McDonagh and G. Tzimiropoulos. Joint face detection and alignment with a deformable hough transform model. In *ECCV*, pages 569–580. Springer, 2016.

[53] A. Paszke, S. Gross, S. Chintala, and G. Chanan. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. *PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration*, 2017.

[54] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *ICAVSBS*, pages 296–301. Ieee, 2009.

[55] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas. A recurrent encoder-decoder network for sequential face alignment. In *ECCV*, pages 38–56. Springer, 2016.

[56] D. Ramanan and X. Zhu. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886. Citeseer, 2012.

[57] E. Richardson, M. Sela, and R. Kimmel. 3d face reconstruction by learning from synthetic data. In *2016 3DV*, pages 460–469. IEEE, 2016.

[58] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, pages 1259–1268, 2017.

[59] S. Romdhani and T. Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *CVPR*, volume 2, pages 986–993. IEEE, 2005.

[60] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*, pages 397–403, 2013.

[61] Z. Sánta and Z. Kato. 3d face alignment without correspondences. In *ECCV*, pages 521–535. Springer, 2016.

[62] J. Saragih and R. Goecke. A nonlinear discriminative approach to aam fitting. In *ICCV*, pages 1–8. IEEE, 2007.

[63] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *IJCV*, 91(2):200–215, 2011.

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

[65] L. Tran and X. Liu. Nonlinear 3d face morphable model. In *CVPR*, pages 7346–7355, 2018.

[66] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *CVPR*, pages 5163–5172, 2017.

[67] G. Tzimiropoulos and M. Pantic. Optimization problems for fast aam fitting in-the-wild. In *ICCV*, pages 593–600, 2013.

[68] R. Yu, S. Saito, H. Li, D. Ceylan, and H. Li. Learning dense facial correspondences in unconstrained images. In *ICCV*, pages 4723–4732, 2017.

[69] 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.

[70] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *CVPR*, pages 787–796, 2015.

[71] A. D. Bagdanov, A. Del Bimbo, and I. Masi. The florence 2d/3d hybrid face dataset. In *J-HGBUW*, pages 79–80. ACM, 2011.
[72] A. Tewari, M. Zollhofer, H. Kim, P. Garrido, F. Bernard, et al. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV*, pages 1274–1283, 2017.
[73] A. Bas, P. Huber, W. A. Smith, M. Awais, and J. Kittler. 3d morphable models as spatial transformer networks. In *ICCV*, pages 904–912, 2017.
[74] T. Hassner. Viewing real-world faces in 3d. In *ICCV*, pages 3607–3614, 2013.
[75] I. Kemelmacher-Shlizerman and R. Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE TPAMI*, 33(2):394–405, 2010.
[76] L. Gu and T. Kanade. 3d alignment of face in a single image. In *CVPR*, volume 1, pages 1305–1312. IEEE, 2006.
[77] C. Bhagavatula, C. Zhu, K. Luu, and M. Savvides. Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In *ICCV*, pages 3980–3989, 2017.
[78] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *CVPR*, pages 1078–1085. IEEE, 2010.
[79] X. Xiong and F. De la Torre. Global supervised descent method. In *CVPR*, pages 2664–2673, 2015.
[80] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, pages 4998–5006, 2015.
[81] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017.
[82] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014.

Xiaoguang Tu is a Ph.D. candidate with the School of Information and Communication Engineering at University of Electronic Science and Technology of China (UESTC). He was a visiting scholar at Learning and Vision Lab, National University of Singapore (NUS) from 2018 to 2020 under the supervision of Dr. Jiashi Feng. His research interests include convex optimization, computer vision and deep learning.

Jian Zhao received the B.S. degree from Beihang University in 2012, the Masters degree from the National University of Defense Technology in 2014, and the Ph.D. degree from National University of Singapore in 2019. He is currently an Assistant Professor with Institute of North Electronic Equipment, Beijing, China, and the Rhino-Bird Visiting Scholar with the Tencent AI Lab, Shenzhen, China. His main research interests include Machine Learning, Pattern Recognition and Computer Vision.

Mei Xie is a professor with School of Electronic Engineering at University of Electronic Science and Technology of China (UESTC). She received the Ph.D. degree in signal and information processing (SIP) from UESTC in 1997. Between 1997 and 1999, she studied in University of HongKong and University of Texas for the postdoctoral degree, respectively. Her research interests include pattern recognition, computer vision and artificial intelligence.

Zihang Jiang is a Ph.D. candidate with Department of Electrical and Computer Engineering at National University of Singapore. He received the B.S. degree in Mathematics from University of Science and Technology of China in 2019. His research interests include computer vision, artificial intelligence and 3D vision.

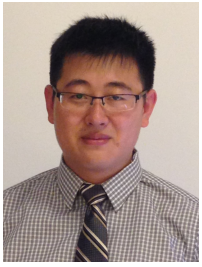
Akshaya Balamurugan is a Masters graduate from Institute of Systems Science at National University of Singapore (NUS). Her research interests include deep learning, computer vision and incremental learning. She is currently an AI scientist at Pensees Ptd Ltd.

Yao Luo is a Masters graduate with School of Information and Communication Engineering at University of Electronic Science and Technology of China (UESTC). He received the master's degree in Electronics and Communication Engineering from UESTC in 2020. His research interests include digital image processing and computer vision.

Yang Zhao received his B.S. degree in National University of Defense Technology, Hunan, in 2014, and received his M.S. degree in computer science and technology at National University of Defense Technology in 2016, and now he is a Ph.D. candidate at National University of Defense Technology. His research interests include machine learning and computer vision.

Lingxiao He is a research scientist in JD AI Research. He received the B.E degree in information Engineer from the Chengdu University of Technology (CDUT), the Ph.D. degree in Computer Sciences from Institute of automation, Chinese Academy of Sciences (CASIA) in 2014, 2019, respectively. He visits Learning and Vision Lab, National University of Singapore (NUS) from September 2018 to May 2019. Since August 2019, Dr. His research areas include biometric, pattern recognition and computer vision.

Zheng Ma is a professor with the School of Communication and Information Engineering at University of Electronic Science and Technology of China (UESTC). His research interests include convex optimization, computer vision and image processing.



Jiashi Feng is currently an Assistant Professor in the Department of Electrical and Computer Engineering at National University of Singapore. He received his Ph.D. from National University of Singapore in 2014. Before joining NUS as a faculty, he was a postdoc research follow at UC Berkeley. Dr. Feng's research areas include computer vision and machine learning. In particular, he is interested in object recognition, detection, segmentation, robust learning and deep learning.